ETHICS OF ARTIFICIAL INTELLIGENCE

EDITED BY S. MATTHEW LIAO

Ethics of Artificial Intelligence

Ethics of Artificial Intelligence

EDITED BY S. MATTHEW LIAO

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

Library of Congress Cataloging-in-Publication Data
Names: Liao, S. Matthew, editor.
Title: Ethics of artificial intelligence / edited by S. Matthew Liao.
Description: New York, NY, United States of America : Oxford University Publication, 2020. | Includes bibliographical references and index.
Identifiers: LCCN 2020004473 (print) | LCCN 2020004474 (ebook) | ISBN 9780190905040 (paperback) | ISBN 9780190905033 (hardcover) | ISBN 9780190905064 (epub)
Subjects: LCSH: Artificial intelligence—Moral and ethical aspects.
Classification: LCC Q334.7 .E84 2020 (print) | LCC Q334.7 (ebook) | DDC 174/.90063—dc23
LC record available at https://lccn.loc.gov/2020004473
LC ebook record available at https://lccn.loc.gov/2020004474

1 3 5 7 9 8 6 4 2

Paperback printed by LSC Communications, United States of America
Hardback printed by Bridgeport National Bindery, Inc., United States of America

For Wibke, Caitlin, and Connor

Acknowledgments

This volume emerged in part from a conference in October 2016 on the ethics of artificial intelligence at New York University hosted by the NYU Center for Mind, Brain and Consciousness and the NYU Center for Bioethics. I would like to thank my co-organizers, Ned Block and David Chalmers, for making the conference such a tremendous success. Thanks are also due to Jonathan Simon, Cassandra Coste, and Leigh Bond, who provided logistical and organization support for all facets of the conference. Tom Carew, the former Dean of the Faculty of Arts and Science at NYU, deserves special thanks for providing the opening remarks at the conference. I would also like to express my appreciation to all the speakers and panelists for their insightful talks and presentations: Peter Asaro, John Basl, Nick Bostrom, Meia Chita-Tegmark, Kate Devlin, Vasant Dhar, Virginia Dignum, Mara Garza, Daniel Kahneman, Adam Kolber, Yann LeCun, Gary Marcus, Steve Petersen, Francesca Rossi, Stuart Russell, Ronald Sandler, Jürgen Schmidhuber, Susan Schneider, Eric Schwitzgebel, Frans Svensson, Jaan Tallinn, Max Tegmark, Wendell Wallach, Stephen Wolfram, and Eliezer Yudkowsky.

I am particularly grateful to David Chalmers, who provided intellectual support and guidance throughout this process. The thirty contributors to this volume merit a special gratitude for their excellent intellectual work. I would like to thank Peter Ohlin, my editor at Oxford University Press, for suggesting the idea of producing this volume and for his enthusiasm and encouragement throughout its production.

Colin Allen, David Chalmers, Mala Chatterjee, Felipe De Brigard, Sarah Gokhale, Robin Hanson, Athmeya Jayaram, Ryan Jenkins, Andrew Lee, Robert Long, Neil McArthur, Rune Nyrup, Huw Price, Duncan Purves, Jonathan Simon, Daniel Viehoff, and Roman V. Yampolskiy read and provided astute comments on various chapters in this book, for which I am very grateful. I have also been helped greatly by Sarah Gokhale, Yulia Gamper, and Nicholas Tilmes, who provided invaluable research assistance.

Finally, I would like to thank my family, Wibke, Caitlin, and Connor, for their love and support as I worked to complete this book.

Contributors

Peter Asaro is an Associate Professor and the Director of the Graduate Program in Media Studies at The New School, and an Affiliate Scholar of Stanford Law School's Center for Internet and Society. He is also the Co-founder and Vice-Chair of the International Committee for Robot Arms Control, and a founding member of the Campaign to Stop Killer Robots.

Jean-François Bonnefon is a Research Director at the French Centre National de la Recherche Scientifique and is affiliated with the Toulouse School of Economics, the Toulouse School of Management, and the Institute for Advanced Study in Toulouse. He is the Moral AI Chair at the Artificial and Natural Intelligence Toulouse Institute.

Nick Bostrom is a Professor at Oxford University, where he leads the Future of Humanity Institute as its founding director. He is the author of over 200 publications, including *Anthropic Bias, Global Catastrophic Risks, Human Enhancement*, and *Superintelligence: Paths, Dangers, Strategies*, a New York Times bestseller which helped spark a global conversation about artificial intelligence.

Andrew Critch is a Research Scientist at the Center for Human-Compatible AI in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. He was a Research Fellow at the Machine Intelligence Research Institute.

Allan Dafoe is an Associate Professor and Senior Research Fellow in the International Politics of Artificial Intelligence at the University of Oxford. He is also the Director of the Centre for the Governance of AI at University of Oxford's Future of Humanity Institute.

Kate Devlin is a Senior Lecturer (Associate Professor) in Social and Cultural Artificial Intelligence at King's College London. She is the author of *Turned On: Science, Sex and Robots*, which examines the ethical and social implications of technology and intimacy.

Carrick Flynn is a Research Fellow at Georgetown University's Center for Security and Emerging Technology focused on national security, technology law, and AI policy. He was the founding Assistant Director of the Center for the Governance of AI at the University of Oxford.

Mara Garza is a former philosophy PhD student at the University of California, Riverside and is currently an artist.

Hanna Gunn is an Assistant Professor in Cognitive and Information Sciences at the University of California, Merced. Her research focuses on themes in social epistemology including agency, community, new media, and the Internet.

Aaron James is a Professor of Philosophy at the University of California, Irvine. He is the author of several books, including *Fairness in Practice: A Social Contract for a Global Economy*.

Frances M. Kamm is the Henry Rutgers University Professor of Philosophy and Distinguished Professor of Philosophy at Rutgers University. She is the author of numerous works in normative ethical theory and applied ethics, including *The Trolley Problem Mysteries*.

Patrick LaVictoire is a machine learning engineer, most recently for Lyft. After a postdoctoral position at the University of Wisconsin, he worked on AI alignment theory at the Machine Intelligence Research Institute.

S. Matthew Liao is the Arthur Zitrin Chair of Bioethics, the Director of the Center for Bioethics, and an Affiliated Professor in the Department of Philosophy at New York University. He is the author or editor of numerous books including *The Right to Be Loved* and *Moral Brains: The Neuroscience of Morality*.

Andrea Loreggia is a Research Associate at the European University Institute. His research interests in artificial intelligence span from knowledge representation to deep learning.

Nicholas Mattei is an Assistant Professor of Computer Science at Tulane University. His research focuses on the theory and practice of artificial intelligence, machine learning, data science, and the impact of these technologies on society.

Cathy O'Neil is the author of *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. She is the founder of O'Neil Risk Consulting & Algorithmic Auditing.

Steve Petersen is an Associate Professor of Philosophy at Niagara University. In addition to the ethics of AI, he works on algorithmic approaches to epistemology.

Iyad Rahwan is a Director of the Max Planck Institute for Human Development, where he founded and directs the Center for Humans & Machines. He is also an Associate Professor of Media Arts & Sciences at the Massachusetts Institute of Technology

Peter Railton is the Gregory S. Kavka Distinguished University Professor and John Stephenson Perrin Professor of Philosophy at the University of Michigan, where he specializes in ethics and philosophy of science. He is also a Fellow of the American Academy of Arts and Sciences.

Francesca Rossi is based at the T.J. Watson IBM Research Lab in New York. Before joining IBM, she was a Professor of Computer Science at the University of Padova.

Stuart Russell is a Professor of Computer Science at the University of California, Berkeley, an Honorary Fellow of Wadham College, Oxford, and an Andrew Carnegie Fellow. He co-authored (with Peter Norvig) *Artificial Intelligence: A Modern Approach*, the standard text in the field. He has been active in the arms control community for nuclear and autonomous weapons.

Susan Schneider is the NASA/Baruch Blumberg Chair of Astrobiology and Technological Innovation at NASA and the Library of Congress and the Director of the AI, Mind and Society Group at the University of Connecticut. She writes about the nature of the self and mind, from the perspectives of philosophy, AI, cognitive science and astrobiology.

Eric Schwitzgebel is a Professor of Philosophy at University of California, Riverside, and a member of its program in Speculative Fiction and Cultures of Science. His most recent book is *A Theory of Jerks and Other Philosophical Misadventures*.

Azim Shariff is an Associate Professor of Psychology at the University of British Columbia where he directs the Centre for Applied Moral Psychology. His research addresses ethical issues related to religion, free will, economics, and technological trends.

Jessica Taylor is a technical philosopher who has researched topics including logical probability, decision theory, ontology, and distributed consensus algorithms. She works at the Median Group, and she was a Research Fellow at the Machine Intelligence Research Institute.

Shannon Vallor is the Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence at the Edinburgh Futures Institute at the University of Edinburgh, where she is also appointed in Philosophy. She is the author of *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*.

K. Brent Venable is a Senior Research Scientist at the Florida Institute of Human and Machine Cognition in joint appointment with University of West Florida, where she is a Professor of Computer Science. She has co-authored two books and published over 100 academic papers.

Wendell Wallach chaired the Technology and Ethics Research Group for the past eleven years at Yale University's Interdisciplinary Center for Bioethics. His latest book, a primer on emerging technologies, is entitled, *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control,* and he co-authored (with Colin Allen) *Moral Machines: Teaching Robots Right from Wrong*.

Stephen Wolfram is the creator of Mathematica, Wolfram|Alpha and the Wolfram Language; the author of *A New Kind of Science*; and the founder and CEO of Wolfram Research. Over the course of nearly four decades, he has pioneered the development and application of computational thinking and has been responsible for many discoveries, inventions, and innovations in science, technology and business.

Eliezer Yudkowsky has been a Research Fellow at the Machine Intelligence Research Institute since 2000, working on decision theory and on AI alignment.

A Short Introduction to the Ethics of Artificial Intelligence

S. Matthew Liao

I.1. Overview

Artificial intelligence (AI) is progressing rapidly. AI can now recognize objects in images and videos; transcribe speech; translate between languages; beat humans at *Jeopardy*,[1] at Go,[2] and at poker;[3] paint in the style of van Gogh;[4] write Beatles-like music;[5] help prepare legal documents; trade stocks; drive cars; fly drones; write its own encryption language;[6] identify cancer in tissues;[7] and solve the quantum state of many particles at once.[8] In the coming years, it is expected that AI will reach and exceed human performance on many more, and increasingly complex, tasks.

As AI technologies continue to advance, questions about the ethics of AI become more pressing than ever. Complex ethical issues surround current and near-future AI systems. For instance, leading technology companies are building self-driving cars, which promise to increase personal mobility for elderly and disabled people and to save lives by reducing driver error; however, in an emergency, should a self-driving car prioritize the lives of the passengers or the lives of pedestrians?[9]

Many countries are developing autonomous weapon systems capable of identifying and attacking a target without human intervention. Autonomous weapon systems offer the potential to decrease risks to military personnel and civilians by being better than stressed-out soldiers in the heat of a battle at distinguishing civilians from combatants and at making reasonable trade-offs between military gains and risk or harm to civilians. But should we as a society give machines this kind of moral decision-making power? Who is at fault if an autonomous weapon system attacks a hospital or a school? Does the exercise of deadly force always require "meaningful human control" to be legitimate?[10]

AI has been found to be particularly useful for revealing otherwise unrecognizable patterns in complex processes. As a result, police departments are investigating ways to use AI to identify likely criminals among the general population.[11] Likewise, some judges and prison officials are interested in using AI to

S. Matthew Liao, *A Short Introduction to the Ethics of Artificial Intelligence* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0001

develop "risk prediction" tools to assist with decisions on criminal sentencing, bail hearings, and parole.[12] However, among other things, machine learning requires good data because incomplete or unrepresentative data can exacerbate problems of bias.[13] At the same time, advanced AI systems tend to process large volumes of data, and their inner workings tend not to be transparent. Given this, how can we create AI systems that are fair and that do not inadvertently produce biased results?

At the same time, Japan, Italy, and some other countries are looking into using robots to care for and provide companionship to their elderly population.[14] Many companies are racing to build sex robots with sophisticated AI.[15] What ethical issues are raised by companion robots and sex robots? Will humans be able to marry robots?[16] How will this impact our relationships with other humans?

It is estimated that 47% of American jobs could be lost to AI and automation in the next twenty years.[17] How should we help people adjust to this level of unemployment? Should taxes be levied on robots that replace human workers, as Bill Gates has suggested?[18] Should we offer everyone a universal basic income?[19]

Current AI is what is known as narrow AI[20] because it is designed to perform a narrowly defined task such as driving a car or identifying a hostile target. In the long term, a number of AI researchers hope to create artificial general intelligence (AGI), which would be capable of performing any intellectual task that a human being can.[21] On one understanding, such AI, sometimes referred to as strong AI, would be capable of actual thought and reasoning and would possess sentience and consciousness.[22] Some writers speculate that once a sufficiently intelligent AI is developed, it could develop even more intelligent systems, which in turn could develop systems with even greater intelligence, resulting in an "intelligence explosion" or "singularity," whereby superintelligent machines would come to possess capacities that greatly exceed human capacities.[23]

The prospect of superintelligent AIs raises at least two kinds of ethical issues. One issue pertains to the impact of these AIs on humans. In particular, can we prevent them from causing harm to humans or even human extinction? To illustrate, consider an example from Nick Bostrom in which a machine designed to make as many paper clips as possible becomes incredibly intelligent.[24] Given its goal, the machine could decide to convert everything in the universe, including humans, to paper clips. Is it possible to shape the development of AI now and align the eventual values and goals of superintelligent AIs with those of humans so that these AIs will not end up harming or destroying humanity?[25] A second issue is how we should treat these AIs. Would such AIs be conscious? Would they have the moral status that humans have, that is, would they be rightsholders?[26] What kind of rights and responsibilities would they have?[27] Could they have moral status greater than that of humans?

In recent years, the ethical implications of near-term and long-term AI have received considerable attention in both the popular media and academia.[28] In this introduction, I aim to outline some of the key issues in the study of the ethics of AI, identify some of the core claims that have been made, and propose some ways of taking these discussions further. With respect to near-term AI, it will be useful to distinguish between (a) ethical issues that arise because of limitations to current machine learning systems, what might be called "vulnerabilities in machine learning," and (b) ethical issues that arise because current machine learning systems may be working too well and humans can be vulnerable in the presence of or when interacting with these intelligent systems, what might be called "human vulnerabilities." The chapters in this volume will then continue this discussion by offering a variety of new perspectives on the ethics of AI.

I.2. Key Concepts in Machine Learning

To begin, let me say something about what AI is.[29] There is no agreed-upon definition of AI. In 1956 at a conference in Dartmouth, John McCarthy coined the term "artificial intelligence" and defined it as "the science and engineering of making intelligent machines."[30] Stuart Russell and Peter Norvig, authors of one of the most popular textbooks on AI, propose that there are four ways of defining AI: as (a) acting humanly, (b) thinking humanly, (c) thinking rationally, and (d) acting rationally. They are interested in AI concerned with (d), that is, rational action.[31] For our purpose, we can broadly understand AI as getting machines to do things that require cognitive functions such as thinking, learning, and problem-solving when done in intelligent beings such as humans.

On this understanding, AI can take different forms.[32] One form is symbolic AI, or good-old-fashioned artificial intelligence (GOFAI), which dominated the field of AI research from the 1950s to the 1980s.[33] Symbolic AI attempts to represent cognitive functions such as thinking, learning, and problem-solving through symbolic reasoning and logic. In particular, these systems use a series of explicitly programmed if-then rules and statements to establish the relations between inputs and outputs. Examples of symbolic AI include rules engines such as expert systems (where the rule set is a representation of an expert's knowledge) and knowledge graphs (where a database stores information in a graphical format). A limitation of symbolic AI is that it is difficult to revise the rules once they are encoded into such a system.

Another form of AI is machine learning, which uses algorithms to learn from data without being explicitly programmed. Within machine learning, one can distinguish between supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an algorithm aims to learn a function that

best approximates the relationship between input and output in the data. To do so, the algorithm is trained on a training data set in which the correct answers for certain data are known and the data are labeled accordingly. In this way, the algorithm can use the labeled information to learn the relationship between inputs and outputs. Once the algorithm is properly trained, it is then able to apply what it has learned to predict the correct answer in different (target) data sets.

There are two main types of supervised learning algorithms: classification and regression.[34] In classification algorithms, the desired output is a discrete label with a finite set of possible outcomes. Cases of binary classification have only two possible outcomes; for example, either something is a car or it is not a car. Cases of multilabel classification have more than two possible outcomes, for instance, text categorization of news articles.[35] In regression algorithms, the desired outputs are continuous.[36]

To give an example of supervised learning, suppose that there is a box containing images of cars and motorcycles. A classification algorithm can be trained by repeatedly manually telling the algorithm which images depict a car and which images depict a motorcycle. Over time, a properly trained algorithm will learn the features that are more likely to make something a car rather than a motorcycle, and vice versa. After that, suppose that there is another box containing new images of cars and motorcycles. The properly trained algorithm should be able to distinguish images of cars from those of motorcycles on its own.

In unsupervised learning, a given data set is not labeled, and the algorithm aims to sort the data on its own. One type of unsupervised learning algorithm is clustering, where the algorithm aims to group data that are more similar to each other than they are to data in other groups.[37] Another type of unsupervised learning is association rule learning, where the algorithm tries to discover rules that describe large portions of the data. To illustrate unsupervised learning, suppose that there is a box with images of cars and motorcycles that have not been labeled or sorted. An unsupervised learning algorithm would attempt to sort and categorize these images based on their similarities and differences.

In reinforcement learning, the algorithm attempts to learn through experience.[38] A reinforcement learning algorithm learns by being rewarded if it succeeds in a task and/or punished if it fails. Through trial and error, the algorithm strives to maximize the long-term reward.

Currently, the driver for many breakthroughs in AI is deep learning. Deep learning uses artificial neural networks, which loosely emulate the activity of neurons in the brain. Simpler versions of such networks existed as far back as the 1950s,[39] but were not taken seriously by the mainstream AI community for decades. Recently, advances in computing power and data storage, coupled with the development of more sophisticated neural networks by researchers such as Geoffrey Hinton,[40] Yann LeCun,[41] and Yoshua Bengio,[42] have unleashed deep learning's potential.[43]

A deep neural network typically has an input layer, an output layer that acts as the final decision-maker, and multiple hidden layers of nodes in between. The "deep" in "deep learning" refers to the number of hidden layers of nodes in a neural network. Each layer is responsible for performing a specific function that contributes to solving the problem at hand. Each node has weights associated with each of its connections. Adjusting the weights on these connections causes a node to produce certain outputs, which are then fed through the following layers.

Deep learning uses these layers of nodes to detect increasingly abstract features of a data set that capture the most information while minimizing losses in accuracy. To do this, it uses the method of back-propagation,[44] which takes the error between the expected result and the actual result of a neural network and adjusts the neural network's weights in the direction of less error. In doing so, the entire network progressively gets better at predicting the correct answer.

As an example, consider again a classification task involving images of cars and motorcycles. The input layer of a deep neural network might take in external data such as pixels in an image. It would then feed this information forward to some or all of the connections in the next layer. Each connection in that layer would integrate the inputs from the first layer and pass the results forward to subsequent layers. Eventually the output layer would offer a prediction, for example, that an image is a car or a motorcycle.

Different types of neural networks are used for different purposes. For example, recurrent neural networks (RNNs),[45] which have a sort of built-in memory, are particularly useful for solving temporal data problems such as predicting sequences of words. Convolution neural networks (CNNs),[46] which are good at recognizing patterns across space, are particularly useful for image classifications and computer vision tasks. Deep learning can be combined with other machine learning methods such as reinforcement learning to create deep reinforcement learning algorithms. A deep reinforcement learning algorithm was used to beat the world champion Lee Sedol at the game of Go.[47]

I.3. Vulnerabilities in Machine Learning

As impressive as machine learning is, it also suffers certain limitations. As we shall now see, these limitations can give rise to a host of ethical issues.

I.3.1. Machine Learning Is Data Hungry

First, machine learning needs a lot of data to work well. For example, supervised learning algorithms can fine-tune themselves and achieve great predictive power

when they have access to a vast amount of data. Consequently, this incentivizes companies and organizations to harvest or buy data, including sensitive, personal data, even when doing so might involve violating an individual's right to privacy. For example, Cambridge Analytica collected millions of Facebook users' data without their knowledge for political purposes.[48] A borderline case might be when the drug maker GlaxoSmithKline bought the exclusive rights to mine the genetic data of customers of the DNA testing service 23andMe for drug discovery.[49]

I.3.2. Garbage In/Garbage Out

Second, machine learning is only as good as the data from which it learns. If a machine learning algorithm is trained on inadequate or inaccurate data, then the algorithm will make bad predictions even if it is itself well designed. For instance, in 2015 Google Photo autolabeled Jacky Alcine, a black software developer, and his friend as "gorillas"[50] because, in all likelihood, the data used to train the algorithm did not include sufficient images of people from different racial and ethnic backgrounds.[51]

I.3.3. Faulty Algorithms

Third, even if a machine learning algorithm receives adequate and accurate data, if the algorithm itself is bad, it will also make bad predictions. For instance, a bad machine learning algorithm may identify a pattern even if there isn't one, a problem known as "overfitting,"[52] or may fail to identify a pattern even when there is one, a problem known as "underfitting."[53] A machine learning algorithm may also give too much or too little weight to certain features or fail to include certain relevant features altogether. Faulty algorithms can have serious ethical implications. For example, in 2016 the Arkansas Department of Human Services began to use an algorithmic tool developed by interRAI to determine how many hours of home care some people with disabilities should receive.[54] The department implemented the algorithm's recommendation to reduce drastically the number of home care hours for many beneficiaries, which caused several people to be hospitalized. After a lawsuit was filed, an investigation revealed that, among other things, the algorithm had incorrectly coded conditions such as cerebral palsy and had not accounted for conditions such as diabetes, which led it to recommend reduced home care hours for hundreds of people. Ultimately a judge ruled that the department had insufficiently implemented the interRAI algorithm and ordered that its use be terminated.

I.3.4. Deep Learning Is a Black Box

Fourth, deep learning is a black box that raises issues such as interpretability, explainability, and trust.[55] Deep learning is impenetrable even to its programmers because it typically employs thousands or millions of connections that interact with one another in complex ways. As a result, it is difficult to interpret how these connections are interacting with each other and why they make certain predictions. The issue of explainability arises because humans often need to know how a decision is reached. However, deep learning announces its prediction without explaining (in human terms) how it arrived at that prediction. To see why this could be a problem, consider the following example. Suppose that a deep learning algorithm predicts that there is a 74% chance that Kay will commit another crime in the future, and a judge denies Kay parole on this basis. The deep learning algorithm does not, for example, say, "There is a 74% chance that Kay will commit a crime in the future because she has committed such crimes in the past, and the last time Kay was on parole, she recommitted a crime within two weeks." Without such an explanation, the judge would not be able to explain, and justify, to Kay why she was denied parole. Beyond explainability, this also raises the issue of trust in the deep learning system, since we do not know whether it makes its predictions on reasonable and reliable grounds. For high-stakes decisions such as those concerning parole, not being able to trust the deep learning system is especially problematic.

Are there ways to address or mitigate deep learning's black box problem? Some AI researchers are currently exploring technical fixes such as "interpretable machine learning."[56] One interpretable machine learning method involves adding an additional layer to deep learning models after the hidden layer(s) of nodes and before the output.[57] The added layer would provide information such as which features were the most important for arriving at a particular prediction, which features could have had an even greater impact on the prediction, how each feature in the data bears on a particular prediction, and how each feature would affect different possible predictions.[58] The hope is that this information would make the deep learning system more interpretable.

While interpretable machine learning is a promising idea, there are reasons to question whether it can alleviate the problems of interpretability, explainability, and trust in deep learning. One concern is that, since it is placed outside the black box, the additional layer provides a post hoc explanation of the black box after the deep learning system has already made its predictions. One might wonder whether such post hoc explanations can identify the actual reasons why the deep learning system gave the predictions that it did. The concern can be put in the form of a dilemma: either the predictions are *based on* these post hoc explanations or they are not. If the predictions are not based on these post

hoc explanations, what is their value? They might just be a post hoc rationalization that does not correspond to how the black box arrived at its predictions. Of course, post hoc explanations could be useful in some contexts even if the original process was not based on them, as long as they adequately captured some aspects of the process.[59] However, given the black box nature of deep learning, it is unclear how we could know whether these post hoc explanations (adequately) captured some of what was going on in the process. Suppose instead that the predictions are based on the post hoc explanations. If so, it should be possible to design a new model using just these post hoc explanations. But if this is the case, it implies that the black box is dispensable. Indeed it suggests a way to test the value of these interpretable machine learning systems. If the black box remained indispensable for making predictions, this would seem to suggest that the post hoc explanations do not completely explain why a black box gave the predictions that it did.

Other people have attempted to address the black box problem by arguing that the importance of interpretability and explainability may be overstated.[60] According to this line of thought, there is a trade-off between accuracy and explainability in deep learning. If a deep learning system can make accurate predictions, so the thought goes, it may not matter in certain cases if it is not interpretable and explainable. For instance, consider medicine. One would think that interpretability and explainability are especially important in medicine, given its high-stakes nature. However, some argue that clinicians often prescribe medications without fully understanding why they work.[61] For example, clinicians frequently prescribe aspirin as an analgesic and lithium as a mood stabilizer despite persistent uncertainty about the mechanisms through which they work.[62]

There are reasons to be skeptical of this particular argument. While it may be correct that we do not fully understand how some medications work, we do have some ideas regarding the causal mechanisms through which they work. For instance, people knew that something from willow trees *causes* fevers and pain to be reduced, even if they did not know about salicylic acid, an active ingredient in aspirin.[63] This is distinct from a deep learning system that works through associations and is, at least for now, unable to track causal relations. Likewise, it is true that for a long time we did not know exactly how lithium stabilized an individual's mood. (The current hypothesis is that lithium moderates glutamate levels in the brain.)[64] Still, we know that lithium causes moods to be stabilized. Again, we cannot say the same about a deep learning system that cannot track causal relations.[65]

It might be asked why it matters whether a deep learning system can or cannot track causal relations. To answer this question, it is worth noting that deep learning is vulnerable to certain kinds of adversarial attacks, which are inputs

that are designed to cause a machine learning model to make a mistake.[66] For instance, deep neural networks are vulnerable to the so-called one-pixel attacks.[67] In one study, by changing just one pixel in an image, researchers were able to get a deep learning algorithm to classify an image of a car as a dog. The researchers found that one-pixel attacks are successful on nearly three-quarters of standard training images and that altering more pixels made this type of attack even more effective.[68] In another study, researchers modified 0.04% of the pixels in an image, that is, about four hundred pixels out of a million. These changes were imperceptible to the human eye. Nevertheless the deep neural network classified a panda as a gibbon "with 99.3% confidence."[69] Recently researchers found that unmodified real-world images can also be used in adversarial attacks.[70] The fact that deep learning networks are vulnerable to these types of attacks suggests that these networks are not learning "real" features of the world such as causal relations or what a macro-level object like a panda really is; instead these deep learning networks are learning only superficial features. For our purpose, if a deep learning network can be tricked in these ways, issues of interpretability, explainability, and trust remain highly relevant, especially in high-stakes domains such as medicine and law where human beings could be harmed.

I.3.5. Machine Learning Is Weak AI

Fifth, in addition to being narrow AI, current machine learning systems are also weak AI in that they do not have self-awareness or consciousness and they cannot think for themselves.[71] As we have seen, they lack understanding of "real" features of the world such as causal relations. Importantly, they also lack a moral sense, that is, the capacity to assess and determine whether an action is right or wrong. Yet machine learning systems are being deployed in situations in which they may have to make moral decisions without human oversight. Can morality and ethical decision-making be built into such AI? If so, how? A whole literature called "machine ethics" or "machine morality" is devoted to addressing these questions.[72] Owing to space, I will not attempt to give a detailed overview of that literature, but here are some main takeaways.

Some people have proposed designing machines that behave ethically by building moral rules and principles into them using overarching ethical theories such as deontology and utilitarianism. For instance, one might create a utilitarian machine that would aim to secure the greatest good for the greatest number. Or one might create a deontological machine that would follow Kant's categorical imperative.[73] A problem with this top-down approach to machine ethics is that it can be difficult to know when to apply a moral rule or principle and when the rule or principle has been satisfied. For instance, it is a well-known

problem for utilitarianism that it is difficult to calculate when acting would secure the greatest good for the greatest number. This calculation problem would also apply to machine learning systems under consideration. A more pressing issue is that not everyone believes that deontology and utilitarianism are correct moral theories. Given this, there are questions about whether it is appropriate to build such moral theories into systems that are likely to affect everyone.

Others have proposed building virtue ethics into machines.[74] According to this line of thought, one should model a machine on what a virtuous agent would do in a particular situation. A concern about this approach is that it is not altogether clear who is a virtuous agent and what a virtuous agent would do in a particular situation. For instance, would Mother Teresa be a virtuous agent? Suppose that she is. How would one know what Mother Teresa would do in a particular situation? Also, some people may not regard Mother Teresa as a moral exemplar. Should we instead let each person decide who the virtuous agent should be? What if someone's idea of a virtuous agent is a racist, malevolent dictator?

Instead of overarching theories, perhaps we could use case-driven approaches from moral philosophy to build ethics into machines. Since the famous trolley dilemmas are often mentioned in this context, it is worthwhile providing some background on them. Philosophers initially used trolley dilemmas to question utilitarianism.[75] Consider the following two cases.

Sidetrack: A runaway trolley is headed toward five people who will be killed. You can hit a switch that will turn the trolley onto a sidetrack where another person sits, thereby killing him instead of the five.

Footbridge: As before, a runaway trolley is threatening to kill five people. You are standing next to a large man on a footbridge and you can push the large man off the bridge. The large man will die, but his body will stop the trolley from reaching the five people, thereby saving the five.

In both cases, the choice is between killing one person and letting five others die. It seems that utilitarianism would say that the actions in both cases are morally on a par and that we should kill the one and save the five. However, many people believe that while it is morally permissible to hit the switch in Sidetrack and kill the one, it is impermissible to push the large man in Footbridge to save the five. Suppose that people's judgments about these cases are correct. This would seem to call utilitarianism into question.

It might be asked, how does one explain the difference in people's judgments about Sidetrack and Footbridge, supposing that there is a moral difference between these two cases? According to Judith Jarvis Thomson, who coined the term, the "Trolley Problem" is the problem of explaining why our judgments differ between these two cases.[76] One explanation for the Trolley Problem appeals to the Doctrine

of Double Effect (DDE), which relies on a distinction between intending harm and merely foreseeing harm. According to one interpretation of the DDE, there is a moral constraint on acting with the intention of doing harm, even when the harm is used as a means to a greater good.[77] However, it is permissible to act with the intention of employing neutral or good means to promote a greater good, even though one foresees the same harmful side effects, if (a) the good is proportionate to the harm, and (b) there is no better way to achieve this good. Using the DDE, one can explain the permissibility of hitting the switch in Sidetrack on the ground that you merely foresee the innocent bystander's death but you do not intend him to be hit as a means to saving the five. In contrast, in Footbridge, because it seems that you intend the innocent bystander to be hit by the trolley as a means to stopping the trolley from hitting the five, it is not permissible for you to push him off the footbridge. The DDE is not without its critics.[78] Other philosophers have offered alternative explanations of the Trolley Problem.[79]

For our purpose, it might be thought that one could use trolley-like scenarios to figure out people's judgments about various cases.[80] One might then be able to program people's judgments about those cases into a machine. There is something to this idea. Consider the following:

> **Empty Track:** A runaway trolley is headed toward five people who will be killed. You can hit a switch that will turn the trolley onto an empty sidetrack where there is no one.

In this case, there should be no disagreement that one should hit the switch and turn the trolley toward the empty track. If so, perhaps one could use clear cases such as Empty Track to create some clear ethical boundaries for machines.

Still, this approach also has problems. For one thing, not every case will be as easy as Empty Track. In more difficult cases, people's judgments are likely to differ. Indeed, even with respect to the original trolley dilemma, some people believe that it is impermissible to hit the switch in Sidetrack, while others believe that it is permissible to push the large man off the bridge in Footbridge.[81] Of course, we could still think that there is a fact of the matter with respect to these cases even though they are difficult. History is replete with examples such as slavery where there were strong disagreements that have been resolved. Even so, at the very least, more philosophical work will need to be done before we can program a machine to act one way or another with respect to these difficult cases. In addition, there are many difficult real-world cases that would need to be resolved before one could build these cases into a machine. Hence, even if we could use the case-based approach to program some easier cases into machines, we are a long way from giving them any sort of complete ethical decision procedure.

Perhaps recognizing that this is a limitation of current AIs, companies typically strive for the less ambitious goal of making sure that their AI systems are safe, akin to having safety protocols for aircrafts. But even meeting this aim can be challenging. Consider an example concerning self-driving cars. To ensure that self-driving cars are safe for everyone involved, manufacturers have to consider, among other things, how the cars should behave when accidents are about to occur. For instance, suppose that a self-driving car's brakes malfunction and the car can either hit five people on the road or swerve off the road and thereby kill the passenger. Suppose that it is not the passenger's fault that the brakes malfunctioned. What should the self-driving car do? Some manufacturers have suggested that the self-driving car should prioritize the passenger in such circumstances.[82] This means that the car should hit the five people instead of endangering the passenger. There is some plausibility to this line of thinking. In ethics, it is commonly held that under ordinary circumstances, we do not have to sacrifice our lives in order to prevent others from being harmed, especially if we are not at fault. Also, practically speaking, most people are unlikely to purchase a car that would sacrifice them for the greater good. Moreover, always prioritizing the passenger seems like a protocol that could be technically implemented in machines in a straightforward manner. The problem, though, is that it is not clear that we should always prioritize the passenger in an accident. Consider the following:

> **Child:** A runaway self-driving car is headed toward a child who will be killed. The self-driving car can swerve slightly to avoid hitting the child. Swerving the car slightly to avoid hitting the child has a low (but not zero) chance of harming the passenger in the car.

In this case, if the passenger should have absolute priority, then the car should not swerve, since swerving introduces some risk to the passenger. Yet it seems that the car should swerve in this case because the risk to the passenger is low and the benefit of not hitting the child is great. This case suggests that it will not be straightforward to devise and implement protocols that would ensure the safety of everyone involved.

Here are some incremental proposals that might help to make self-driving cars safer. First, whatever safety protocols we end up adopting for self-driving cars, it seems that the testing of self-driving cars in the "wild," that is, in actual streets, should be much more regulated. Just as pharmaceutical companies are not permitted to test new drugs by giving them randomly to people on the street, perhaps some oversight body should determine when and where self-driving cars can be tested in consultation with members of the community. Second, given that self-driving cars are weak AIs, and given the difficulty of anticipating every

possible scenario on the road, at least for now we should consider having dedicated lanes for self-driving cars.[83] This would free self-driving cars from having to cope with the unpredictability of human driving, among other things. Third, we might consider equipping self-driving cars with devices that would enable them to communicate with each other so that they could coordinate their actions and reduce the number of accidents. Of course, we would also need to protect such devices from hackers and ensure that they would not undermine users' privacy.

I.4. Human Vulnerabilities

In the previous section, we discussed ethical issues that can arise because current machine learning systems are limited in certain ways. In this section, we shall consider ethical issues that can arise because current machine learning systems may be working too well and humans can be vulnerable in their presence. I shall give four examples of such human vulnerabilities, although there are certainly others. I shall also consider how we should address these vulnerabilities.

First, facial recognition technologies can already detect faces in a crowd with great accuracy.[84] In the near future these technologies will likely be able to track constantly anyone who enters a public space at any hour of the day. On the positive side, these technologies can help police find criminals more quickly and identify missing or kidnapped children. On the negative side, a government could use this technology to monitor its citizens or to profile and discriminate against minorities. For instance, a controversial study from Stanford University allegedly found that a machine learning system could correctly distinguish between gay and straight sexual orientation 81% of the time for men and 74% of the time for women just by examining photos of their faces.[85] A government that criminalizes homosexuality could use such facial recognition technology to identify and discriminate against homosexuals. For our purpose, this is an example where machine learning may be working too well and ethical issues arise because people may be tempted to use it for ill.

Second, we are on the cusp of being able to use machine learning to fabricate videos so realistic that humans cannot tell that they are fake. These so-called "deepfakes" use generative adversarial networks to produce new types of data out of existing data sets.[86] One can use this technique to create videos of a person saying or doing things that he or she has never said or done. For instance, the director Jordan Peele and his brother-in-law, BuzzFeed CEO Jonah Peretti, used this technology to produce a video in which President Barack Obama declared that the villain Killmonger in the film *Black Panther* was "right" about his plan for world domination.[87] While amateur hobbyists might use deepfakes to make

people appear to say or do funny things for entertainment, bad actors could use these digital forgeries to conduct smear campaigns against politicians or private citizens and to spread fake news that erodes trust in our institutions.[88] It should be noted that these deepfakes are not yet good enough to fool us completely. However, there is evidence that digital forgeries need not be very convincing for them to be believed and cause significant damage. For instance, someone manipulated a video of the current U.S. House speaker, Nancy Pelosi, by slowing the speed of the video to 75%.[89] This was enough for the video to go viral and for people to accuse Pelosi of slurring her words. In any case, deepfakes will likely advance to a point where it will be difficult for us to detect whether or not they are fake. For our purpose, deepfakes serve as another example where machine learning is working too well and can be used to exploit our tendency to believe what we see, which is reasonable in ordinary contexts but less so as deepfakes proliferate.

Third, given that robots can perform certain tasks better and faster than humans, do not need sleep, can be duplicated and replaced, and so on, many people are worried that robots and automation will replace a significant portion of current human labor in the near future. A study from McKinsey suggested that by 2030, 30% of human labor could be replaced by automation.[90] Another study from Oxford University predicted that 47% of jobs in the United States will be under threat from intelligent machines in the next two decades.[91] Yet some people believe that while automation will make some jobs obsolete, it will also create new ones.[92] After all, there were similar concerns during the Industrial Revolution about machines taking over human jobs, but as it turned out, while the spinning jenny and the steam engine did displace some workers, they also created many new jobs in textiles and manufacturing. Also, automation is likely to replace tasks that are more repetitive and undesirable, which means that at least in the short term, jobs that are more creative will still require humans.[93] Nevertheless it seems certain that some people will lose their jobs as a result of increased automation and that a subset of these people will not be able to transition to new jobs.

Fourth, as robots become more and more sophisticated, some people have begun to regard them as companions. For instance, in 2018 Akihiko Kondo, a Japanese school administrator, married the hologram of the popular anime character Hatsune Miku.[94] In *Love and Sex with Robots*, David Levy predicts that some people will come to prefer robot companions over humans in the future.[95] Similarly companies are developing robot caregivers for the elderly that can bring drinks to them, remind them to take medication, and play games with them.[96] There is also evidence that some elderly people are becoming attached to their robots.[97] However sophisticated, at least given the current state of machine learning, these robots are still weak AI and are not full agents. As such,

among other things, there is a concern that these are not genuine, reciprocal relationships.

How should we address issues that arise because machine learning systems are working too well and humans are vulnerable in their presence? I would like to suggest that we adopt a *theoretical* human rights framework. To see why, let me first briefly say what human rights are and offer a theoretical account of what human rights we have. I shall then suggest that a theoretical framework enables us to explain why certain rights claims are indeed genuine human rights.

Human rights, as A. J. Simmons states, are "rights possessed by all human beings (at all times and in all places), simply in virtue of their humanity."[98] But which features of humanity ground human rights? Elsewhere I have defended what I call a Fundamental Conditions Approach to human rights, which says that human rights protect the fundamental conditions for pursuing a good life.[99] All too briefly, the fundamental conditions are various goods, capacities, and options that human beings qua human beings need, whatever else they qua individuals might need, in order to pursue certain basic activities. Some basic activities include deep personal relationships with one's partner, friends, parents, children; knowledge of the workings of the world, of oneself, of others; active pleasures such as creative work and play; and passive pleasures such as appreciating beauty. The fundamental goods are resources that human beings qua human beings need in order to sustain themselves corporeally, including food, water, and air. The fundamental capacities are powers and abilities that human beings qua human beings require in order to pursue the basic activities. These capacities include the capacity to think, to be motivated by facts, to know, to choose an act freely (liberty), to appreciate the worth of something, to develop interpersonal relationships, and to have control of the direction of one's life (autonomy). The fundamental options are those social forms and institutions that human beings qua human beings require if they are to be able to exercise their essential capacities to engage in the basic activities. These social forms and institutions include the options to have social interaction, to acquire further knowledge, to evaluate and appreciate things, and to determine the direction of one's life.

The Fundamental Conditions Approach can explain why many of the rights in the Universal Declaration of Human Rights are genuine human rights. For instance, consider the right to life, liberty, and security of person (Article 3). Whatever else human beings (qua individuals) need, they (qua human beings) need life, liberty, and security of person in order to pursue the basic activities. If they are not alive, if they cannot freely choose to act to some degree, or if the security of their person is not guaranteed, then they cannot pursue the basic activities. Given this, on the Fundamental Conditions Approach, human beings would have human rights to life, liberty, and security of person. Or consider the right to freedom of thought, conscience, and religion (Article 18), the right

to freedom of opinion and expression (Article 19), and the right to freedom of peaceful assembly and association (Article 20). As I said earlier, one of the fundamental conditions for pursuing a good life is being able to choose freely to pursue the basic activities. In order to choose freely to pursue the basic activities, one must have freedom of expression, thought, religion, and association. On the Fundamental Conditions Approach, human beings would have human rights to freedom of thought, expression, religion, and association.

A theoretical human rights framework such as the Fundamental Conditions Approach also has the resources to explain why certain claims may not be genuine human rights. This gives such a framework an advantage over approaches that simply assume that all claims listed in international human rights documents are genuine human rights. Consider the right to periodic holidays with pay, which appears in Article 24 of the Universal Declaration of Human Rights. Is there such a human right? On the Fundamental Conditions Approach, the important question to ask is whether paid holidays are a fundamental condition for pursuing a good life. That is, are paid holidays something that human beings (qua human beings) need whatever else they (qua individuals) might need in order to pursue the basic activities? There is no doubt that human beings need some rest and leisure in order to pursue the basic activities. Without time for leisure, human beings would not have sufficient time to pursue the basic activities. Given this, some amount of leisure, in the form of holidays, is a fundamental condition for pursuing a good life. However, it does not seem that paid holidays are a fundamental condition for pursuing a good life, because it seems that human beings could pursue the basic activities even if their holidays were not paid. It might be thought that if holidays were not paid, then some people would not be able to afford to take holidays. But this seems to conflate one's right to certain minimum welfare, which one has, with a right to paid holidays. If one cannot afford to take time off work unless one's holidays are paid, one has a human right to certain minimum welfare assistance. But one does not have a human right to paid holidays because paid holidays are not a fundamental condition for pursuing a good life. Note that while there may not be a human right to paid holidays, this does not mean that there could not be a legal right to paid holidays. It goes without saying that there are other sources of normativity besides human rights (e.g., consideration of justice and/or equality), and some of them may ground social goods such as paid holidays.

There are many reasons why we should adopt a human rights framework in addressing human vulnerabilities that can arise from our interactions with machine learning systems. One reason is that respecting and promoting human rights is compulsory One cannot choose not to do it. A second reason is that on one view, human rights are rights against every able person in appropriate circumstances. For instance, Maurice Cranston says, "To speak of a universal

right is to speak of a universal duty. . . . Indeed, if this universal duty were not imposed, what sense could be made of the concept of a universal human right?"[100] On this view, everyone in appropriate circumstances has a duty to protect and promote everyone's human rights. For our purpose, this means that governments, corporations, and even individual AI researchers are all responsible for being proactive in ensuring that the technologies they are developing and employing not only do not violate human rights but also promote human rights. To give an example, companies or AI researchers might think that as long as a user has signed an informed consent form or an End User License Agreement giving them permission to access the user's personal data, then they can do anything with it as long as they abide by the terms of the agreement. The human rights perspective implies that this may not be so. The companies and the AI researchers are responsible for ensuring that they do not use the user's data in ways that could undermine that user's or some other users' human rights. A real-life example of this may be when Google's employees protested against Google's contract for a US Department of Defense program known as Project Maven, which involved Google helping the US government analyze drone footage using AI.[101] Google subsequently decided not to renew this contract, which could be seen as a case where AI researchers and their company took seriously their responsibility to uphold human rights.[102]

A third reason is that many of the issues mentioned here involve human rights, and the human rights framework enables us to see which values are in conflict. For instance, consider a government's use of facial recognition technologies in public spaces. What is at stake is between a government's interest in law and order and our human rights to privacy, freedom of expression, and freedom of association, and our right against discrimination. In this case, we might ask whether it is necessary and justified for a government to monitor the public 24/7, thereby threatening their citizens' rights to privacy and so on in order to maintain law and order. Here is a reason to think that such mass surveillance is not justified. Suppose that in the future, safe and minimally invasive implantable biometric devices that can track an individual's movements and possibly even his or her thoughts become available. Suppose that mass surveillance were justified for the purpose of maintaining law and order. This would seem to imply that governments would also be justified in requiring citizens to have such implants. If this seems like an overreach of a government's authority, it is also an overreach of a government's authority in our current situation.

Likewise, consider deepfakes. On the one hand, people have a right to freedom of speech and expression, and there is the danger that governmental regulation may result in increased censorship. One might also add that in politics especially, it is generally accepted that people are free to use information out of context as satire in order to criticize politicians. On the other hand, information that is

false, misleading, and intended to deceive the public undermines people's ability to exercise their agency, including their political agency. The 2016 US presidential election demonstrated that a foreign government can spread false information online using social media algorithms and potentially influence voters and election results.[103] The human rights framework reveals the tension between respecting people's right to freedom of expression and making sure that their right to effective agency, including political agency, is not undermined.

Consider automation, which promises to increase economic productivity but also threatens people's livelihood as well as economic inequality. At least in present economic systems, people need to work in order to be paid so that they can meet their fundamental needs and pursue the basic activities that they choose. Some people also derive meaning and a sense of self-worth from their work. However, if and when some people's jobs are eliminated, we need to consider, as a society, how these people can meet their fundamental needs and whether they can obtain their sense of self-worth from other sources. Among other things, this raises the issue of whether there is a right to work and/or a right to welfare assistance. As mentioned at the outset, one proposed solution to automation is a universal basic income, that is, a fixed income that governments provide for everyone.[104] Several countries, including Finland and Canada, have experimented with basic income schemes.[105] Other common ideas are establishing training programs to help people transition to other economic sectors and raising wages while reducing working hours. With respect to the universal basic income schemes, it is not clear that *everyone* should receive such a fixed income. For instance, it is not clear that Jeff Bezos, the CEO of Amazon and currently the wealthiest individual in the world, should also receive a universal basic income. Nevertheless, from the perspective of the Fundamental Conditions Approach, it seems that people should have human rights either to be able to obtain the fundamental goods for themselves or to be provided with such goods in the form of welfare assistance should they not be able to obtain these goods.

Lastly, consider robot companions. It is true that given the current state of machine learning, these robots would be unable to participate in a genuine reciprocal relationship. Still, one might think that fully informed adults have a right to decide with whom they would like to associate. The matter may be more complicated with elderly people. It is known that elderly people are at increased risk of being socially isolated and feeling lonely. Indeed almost half of older women (46%) age seventy-five and older live alone.[106] By the time people reach age eighty-five, 40% will live by themselves.[107] There is ample evidence that social isolation and loneliness in older adults are associated with increased mortality and with other adverse health effects, including dementia, increased risk for hospital readmission, and increased risk of falls.[108] Research also shows that elderly people benefit both mentally and physically from feeling socially and

emotionally connected and involved.[109] However, it is doubtful that current robot caregivers can provide the kind of emotional connection that an elderly person needs. Given this, as a society we need to think of other ways of meeting an elderly person's emotional needs.

Here it is worth mentioning that in recent years, the AI community and various international organizations have put forward several codes of practice and ethical guidelines for the use and deployment of AI. Such efforts include the Future of Life Institute's Asilomar Principles,[110] the Partnership on AI's tenets,[111] and the European Commission's Ethics Guidelines for Trustworthy Artificial Intelligence.[112] As I see it, the human rights framework articulated here is compatible with many of the principles found in these guidelines, and could in fact serve as a ground for some of these principles.

I.5. Long-Term AI Issues

As I said, the prospect of superintelligent AIs raises the issues of (a) how we should treat these AIs and (b) how we can make sure, or at least make it more likely, that these AIs will not treat us badly. I shall discuss how we should treat various kinds of AIs, including superintelligent AIs, in my chapter for this volume, so I shall not discuss this issue here. The issue of how to make it more likely that a superintelligent AI will not harm us or destroy humanity is sometimes called the control problem.[113] Before considering various ways of dealing with the control problem, there are two preliminary matters worth addressing. One matter concerns terminology. In the literature, the terms "artificial general intelligence (AGI)" and "human-level AI" are sometimes used interchangeably.[114] While this usage is not inaccurate per se, it would be better to keep these terms distinct since they often refer to different things. As I see it, "general intelligence" means "being able to perform a wide range of tasks" and is typically contrasted with "narrow intelligence," which means "being able to perform a specific task." For instance, Deepmind's AlphaGo has narrow intelligence in the sense that it can only play Go and do nothing else. If so, "artificial general intelligence" should mean "an AI that is able to perform a wide range of tasks." Whatever human intelligence involves, a case can be made that human intelligence, taken as a whole, involves cognitive, emotional, and moral intelligence.[115] If so, "human-level AI" should mean something like "artificial intelligence that has similar kinds of intelligence as a human being, namely, cognitive, emotional, and moral intelligence." On these definitions, AGI and human-level AI could of course be used interchangeably if and when "being able to perform a wide range of tasks" is taken to mean "having cognitive, emotional, and moral intelligence." However, the term AGI is often used more narrowly. In particular, AGI is often intended to mean

something like "being able to perform a wide range of tasks, *cognitively or rationally speaking*." Indeed, recall that Russell and Norvig are interested in the kind of AI that can act rationally as opposed to humanly. On such an understanding of AGI, AGI and human-level AI would not mean the same thing, since AGI would mean something like "having cognitive but not emotional or moral intelligence." Or consider again the example of a superintelligent machine designed to make as many paper clips as possible and that has the ability to convert everything in the universe, including human beings, into paper clips. This superintelligent machine may have cognitive superintelligence, but it seems that it lacks emotional and moral intelligence, since a morally intelligent being should, among other things, recognize the wrongness of killing human beings in order to create more paper clips.[116] If so, it may be appropriate to say that this superintelligent machine has artificial (super) general intelligence, but it would be inaccurate to say that it has (super) human intelligence. Given that the terms "AGI" and "human-level AI" are often used to mean different things, it seems better to keep them apart. To avoid confusion, it may be helpful to refer to such AIs as "artificial general cognitive intelligence" (AGCI) and "AI that has cognitive, emotional, and moral intelligence on par with human beings" (ACMI), respectively.

Another preliminary matter is that many people believe that superintelligence requires strong AI, that is, something that has all the mental powers of a human being, including (phenomenal) consciousness and understanding, and they doubt that strong AI is possible or probable.[117] Consequently, they do not think that we need to be concerned about the control problem.[118] Three points are worth mentioning here. First, some AI researchers believe that there can be AGI without strong AI.[119] That is, they believe that a machine could have the ability to learn and perform different tasks across different domains even if it did not have consciousness or understanding. Arguably, AGCI need not be strong AI. If these researchers are right, then even if strong AI were not possible, this would not preclude the possibility of AGI/AGCI.

Second, it may be the case that human-level AI requires consciousness and/or understanding. As I have said, ACMI involves having cognitive, emotional, and moral intelligence. It may be the case that for an entity to have moral intelligence, it needs to have consciousness and/or understanding. Indeed, on one view, to have moral intelligence and be a moral agent is to be able to take something as a (moral) reason for action.[120] One might think that an entity can take something as a reason for action only if it can understand why something is a reason for action. Moreover it may be the case that in order to recognize that certain kinds of moral reasons exist, an entity needs to be able to appreciate what it is like to be in a certain state.[121] For instance, to recognize that there is a moral reason not to inflict pain on (nonconsenting) sentient creatures, one may need to be able to appreciate what it is like to feel pain. An entity that does not have consciousness,

understood as having some kind of subjective experience, is, by definition, not able to appreciate what it is like to feel pain. If so, it might be thought that an entity that lacks consciousness would be unable to recognize the moral reason not to cause a sentient creature pain. If ACMI does require consciousness and/or understanding, the possibility of ACMI may very well depend on whether strong AI is possible.

Third, supposing that ACMI requires consciousness and/or understanding, there are reasons to think that ACMI could nevertheless exist. To see this, consider the following scenario.[122]

> **Gradual Substitution**: It is year 2100. You find yourself becoming more and more forgetful. You visit your doctor, who informs you that you have early onset Alzheimer's. Brain imaging shows that some of your brain cells are deteriorating. The doctor tells you that those carbon-based cells can be replaced with functionally equivalent inorganic substitutes, thereby restoring your memory and associated brain functions. You decide to go ahead with the procedure. Sure enough, your memory and brain functions are restored and you feel like your old self again. But a couple of months later, you begin to forget things again. So you go back to your doctor and the doctor informs you that some of your other brain cells have deteriorated. The doctor offers you the option of replacing those cells with inorganic substitutes. Again, you choose to have the procedure. After the procedure, you feel like your old self again. This process continues until gradually all of your carbon-based cells are replaced with inorganic substitutes. At the end of the process, you still act like your old self.

For our purpose, supposing that Gradual Substitution were possible, you will have become an ACMI at the end of the process. Suppose that this is the case and suppose that you had consciousness and understanding before the procedure. Have you retained your consciousness and understanding throughout, and at the end of, this process? There are three possibilities.[123] The first is that at some point during the process, you lose consciousness and understanding. You may still act like your old self, but you are not able to feel like your old self. Another possibility is that during the process, your consciousness and understanding gradually dim and become less and less. A third possibility is that you retain your consciousness and understanding at the end of the process. The first and second possibilities seem less plausible than the third. If so, there are some reasons to think that a human-level AI with consciousness and understanding could exist.

Let us now consider three ways of dealing with the control problem. The first aims to ensure that a superintelligent AI would value and respect humanity. Along this line, some people have suggested that we should build into AI explicit rules against harming humanity. For instance, more than fifty years ago

the science fiction writer Isaac Asimov put forward the famous Three Laws of Robotics:[124]

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Other people have recommended that we should make sure that the values and goals of a superintelligent AI are aligned with human ones.[125] The thought here is that superintelligent AIs do not have to bear ill will toward us in order to harm us inadvertently, if their values are misaligned with ours. For instance, a superintelligent paper clip maximizer need not bear ill will toward us; nevertheless, as long as its goal is to make as many paper clips as possible, the paper clip maximizer could conclude that human bodies could be used to produce even more paper clips. If so, we could become extinct just because the paper clip maximizer's value is misaligned with our values of preserving and protecting humanity.

There are issues with these proposals. For example, what Asimov's stories teach us is that building into an AI explicit laws against harming humanity does not seem to work. Indeed the premise of most of his novels is that the Three Laws of Robotics repeatedly fail to prevent robots from harming humans in various situations.[126] With respect to the idea of value alignment, some central challenges include figuring out which human values should be aligned and how we could get a superintelligent AI to adopt and retain them.[127] However, without discussing these proposals further, there is a more general concern with this approach to the control problem. In particular, if a superintelligent AI is much smarter than we are, then it is likely to develop its own values and make its own decisions. In such a case, it seems unlikely that we would be able to guarantee that the decisions that this superintelligent AI would make would be beneficial or friendly to us. This is so even if such an AI would value us for our own sake. To see this, consider the fact that we may value a cat for its own sake. Nevertheless we may also believe that the cat is expendable if the cat could be used to save human lives. Similarly, even if a superintelligent AI believes that we are valuable for our own sake, it could also believe that we are expendable if we could be used, for example, to protect the lives of other superintelligent AIs.

A second way to deal with the control problem aims to degrade the capacities of a superintelligent AI so that the superintelligent AI is unable to harm us. One idea is the AI-sandboxing method, which would constrain a superintelligent AI by isolating it and making sure that it is unable to act on the external

world without our approval.[128] Another idea involves installing a kill switch into a superintelligent AI so that if it were to engage in activities that endanger humans, we would be able to shut the AI down by hitting the kill switch.[129] A general problem with this approach is that, again, if an AI is much smarter than we are, it will be able find ways to regain or retain its capacities. For instance, a superintelligent AI is likely to have the capacity to be persuasive. If so, it may be able to persuade a human gatekeeper to let it out of its sandbox by promising to solve really difficult problems such as developing a cure for cancer or finding a solution to climate change. Similarly, a superintelligent AI is likely to be able to find ways to make copies of itself, rendering a kill switch ineffective.

A third way of dealing with the control problem is encapsulated in the slogan "If you can't beat them, join them." The idea here is that we could try to become super-smart ourselves in order to keep up with a superintelligent AI. How might we be able to do this? In recent years, advances in the biomedical sciences have led to the development of human enhancement technologies that promise to help people to think better, feel happier, and have increased moral sensibility. There are, for example, various pharmacological means of amplifying and enhancing our cognitive, emotional, and moral capacities. These include Ritalin and Modafinil for improving attention and memory, Prozac for helping people feel better, happier, and more energized, and oxytocin for increasing trust.[130]

In addition to these pharmacological means of human enhancement, there are also brain-computer interface (BCI) technologies, which could further enhance our capacities in more targeted ways.[131] BCIs aim to create a direct communication pathway between a brain and an external device by reading and recording brain activity in order to decode its meaning, and writing to specific regions in a brain to manipulate its activities and functions. Brain signals can be recorded and/or manipulated either noninvasively, partially invasively, or invasively. For instance, electroencephalography is a noninvasive method that involves placing electrodes along the scalp in order to record electrical activity produced by neurons in the brain. Transcranial direct-current stimulation (tDCS) is a noninvasive neuromodulatory technique that delivers a low electric current to the scalp. A partially invasive method is electrocorticography, which places electrodes directly on the surface of the brain in order to record electrical activity from the cerebral cortex. An invasive method of affecting brain function is deep brain stimulation (DBS), which involves inserting a thin electrode through a small opening in the skull into a specific area in the brain; the electrode is then connected by an insulated wire to a battery pack underneath the skin; the battery pack sends electrical pulses via the wire to the brain. At present, noninvasive techniques tend to be safer, while invasive approaches tend to produce better results since they involve placing electrodes closer to the neurons.

BCIs are already being used for therapeutic purposes such as ameliorating the effects of Parkinson's disease, epilepsy, and depression. For example, about 100,000 people around the world today have a DBS implant for these conditions.[132] Researchers are also actively looking into whether people who are paralyzed can use BCIs to control prosthetic limbs and generate speech by thought.[133]

There is evidence that BCIs can be used for enhancement purposes.[134] For instance, numerous studies have found that using noninvasive stimulation such as tDCS in the dorsolateral prefrontal cortex, a brain region responsible for working memory, can improve memory and learning.[135] Likewise studies using invasive techniques such as DBS have found that memory can be improved by stimulating the hippocampus and the entorhinal cortex.[136]

Current BCIs tend to be open-loop systems in that they depend on a user's inputs. For example, with respect to DBS, a user has control over the battery pack and is responsible for deciding when and how much electrical stimulation his or her brain should receive. BCIs are moving toward closed-loop systems, that is, systems that do not require user input and that can (a) read and monitor the brain's activities in real time using neural recording and AI and (b) automatically intervene in these activities through electrical stimulation.[137] Through automated algorithms, closed-loop BCIs promise to be even more effective at monitoring and predicting an individual's next actions. For this reason, a number of people believe that future generations of BCIs could be a way for us to keep up with superintelligent AIs. For instance, Elon Musk founded the company Neuralink precisely in order to look for ways in which humans can stay competitive against superintelligence by merging a human brain with a digital brain. Already, Neuralink has announced that they have created flexible, thin "threads" that are less likely to damage the brain than the materials used in current BCIs.[138]

Setting aside the issues of safety and feasibility, it remains uncertain whether closed-loop BCI systems would solve the control problem. After all, once BCIs become closed-loop systems that run on automated algorithms, we may have little or no control over such devices. In other words, once the digital brain is imbued with AI and becomes radically more intelligent than the biological brain, we may again have a control problem in that the biological brain may no longer be able to control the digital brain.

I.6. The Structure of the Volume

This volume is distinctive because it brings together some of the most prominent AI researchers and academic philosophers and presents some of the most important perspectives on selected topics surrounding the ethics of AI today.

In other words, this volume does not seek to be comprehensive or focus exclusively on topics such as robot ethics, the ethics of big data and privacy, or the existential risk for humanity, each of which is or could be a volume in its own right. Instead it uses broad strokes to highlight some of the central themes in the ethics of AI.

The volume has four parts. Part I presents some of the latest thinking on how we can and should build ethics into AI. To start, Peter Railton echoes our concern that the continuing disagreement over overarching ethical theories such as deontology, utilitarianism, and virtue ethics means that it may not be appropriate to program directly these theories into machines. According to Railton, if we nevertheless want to develop machines that are as ethically trustworthy as ordinary humans, we should consider models that resemble human moral learning. In particular, just as infants learn ethics in part by observing adult behavior, Railton proposes that we could build machines that would also be able to learn ethics by observing human behavior.

As one of the chief proponents of case-driven approaches to moral philosophy, Frances Kamm first points out that many cases that are presented as "Trolley Problem" cases in the AI ethics literature in fact raise moral issues distinct from those raised by standard Trolley Problem cases. Kamm discusses some moral issues raised by self-driving cars, such as the role and responsibility of those who program such cars, the liability of pedestrians and drivers to be harmed by such cars, and whether voluntary passengers of self-driving cars are even more liable to be harmed than pedestrians.

Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan argue that even if ethicists were to agree about which ethical principles should guide a moral algorithm, their work would have little impact if laypersons strongly disagreed with them and decided to opt out of using the algorithm. To avoid such "ethical opt-out," as they call it, Bonnefon, Shariff, and Rahwan argue that the field of moral psychology should be deployed to help illuminate people's preferences about the ways machines should handle ethical trade-offs, so that our building of AI systems is informed by these preferences.

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable offer a computer science perspective on how to build intelligent systems that behave morally. They argue that Conditional Preference networks, which graphically represent conditional and qualitative preference relations, can be used to model, combine, and compare subjective preferences and ethical priorities. In particular, they propose that one can measure the distance between an agent's subjective preference and the ethical principles of the agent's community, and they recommend that if the distance between the two is too great, an agent should be guided toward less-preferred actions that are nonetheless compliant with the ethical principles of the agent's community.

Stephen Wolfram, who has spent the past couple of decades building the Wolfram Language as a computational communication language to provide a bridge between human goals and computational capabilities, offers another computer science perspective on how to build ethical AIs. According to Wolfram, the development of symbolic discourse language makes it possible to build an "AI constitution" that defines how we want AIs to act and what ethics they should follow.

Part II explores in greater detail a number of ethical issues arising out of the near-term use of AI. As noted earlier, many people believe that automation will result in technological mass unemployment. Aaron James is concerned that, among other things, this will have terrible social and political consequences, such as the rise of authoritarianism and the hollowing out of democracy. To avoid these consequences, James advocates and defends what he calls a "precautionary basic income," which is a guaranteed minimum income for everyone that would lower the risk of technological mass unemployment. James explains how his precautionary scheme differs from other basic income schemes.

As the militaries of technologically advanced nations seek to apply increasingly sophisticated AI to weapon technologies, a host of ethical, legal, social, and political questions have arisen. Central among these is whether it is ethical to delegate the decision to use lethal force to an autonomous system that is not under meaningful human control. Further questions arise as to who or what could or should be held responsible when such systems use lethal force improperly. Peter Asaro argues that current autonomous weapons are not legal or moral agents that can be held morally responsible or legally accountable for their choices and actions. Given this, according to Asaro, humans need to maintain control over such weapon systems to ensure that the use of such weapons is morally justified in each and every case.

Many AI algorithms are currently already being used to guide decisions in advertising, credit ratings, sentencing of criminals in the justice system, and more. Cathy O'Neil and Hanna Gunn argue that there is a pressing need to recognize and evaluate the ways that structural racism, sexism, classism, and ableism may be embedded in and amplified by these AI systems. To facilitate more robust ethical reflection in AI development and implementation, O'Neil and Gunn propose an ethical matrix that incorporates the language of data science so that nonethicists, including data scientists, can use this tool to analyze their AI design process.

Sex robots are becoming a commercially viable reality. According to Kate Devlin, the sex robots being developed today have a very specific female-gendered embodiment, which runs the risk of objectifying women. To address

this concern, Devlin proposes that we move away from human-like, human-size dolls toward sex robots that take nonhuman forms.

Part III examines new issues relating to the long-term impact of superintelligence on humanity and how we might be able to align the values of a superintelligent AI with human values. Nick Bostrom, Allan Dafoe, and Carrick Flynn consider what a desirable approach to governance in an era of superintelligent machines could look like, and they identify four policy desiderata that they believe should be given extra weight in long-term AI policy: efficiency, allocation, population, and process.

Stuart Russell provides his latest statement on the need for "provably beneficial AI," which involves training machines to learn underlying human preferences by observing human behavior. Russell discusses the technical challenges involved in building provably beneficial AI and responds to some possible concerns to this approach.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch observe that there are two approaches to AI value alignment, namely, specifying the right kind of objective functions and designing AI systems that avoid unintended consequences and undesirable behavior. They survey eight research areas based around these two approaches including how machine learning systems can be trained to detect cases wherein the classification of test data is highly underdetermined, how these systems can learn to imitate humans who are engaged in complex and difficult tasks, and how these systems can be taught not to manipulate and deceive their human operators.

According to Wendell Wallach and Shannon Vallor, the "value alignment" approach to dealing with superintelligent AIs tends to employ computationally friendly concepts such as utility functions, system goals, agent preferences, and value optimizers, which, they believe, do not have intrinsic ethical significance. Wallach and Vallor propose that human-level AI and superintelligent systems can be assured to be safe and beneficial only if they embody something like virtue or moral character.

Steve Petersen points out that the "value learning" approach to AI safety faces three problems: first, it is unclear how any intelligent system could learn its final values; second, it remains uncertain how one determines the content of a system's values based on its physical or computational structure; third, it remains disputed which values the system should aim to learn. Petersen argues that a "miktotelic" approach, which blends together a complex, learnable final value out of many simpler ones, may provide a way for the value learning approach to address these problems.

Part IV considers how we might be able to determine whether an AI has consciousness and how we should treat AIs that have human-level capacities. Susan

Schneider proposes a provisional framework for investigating artificial consciousness that involves several tests or markers. One test is the AI Consciousness Test, which challenges an AI with a series of increasingly demanding natural-language interactions to see how quickly and readily it can grasp and use concepts based on the internal experiences we associate with consciousness. Another test is based on the Integrated Information Theory and considers whether a machine has a high level of "integrated information." Third is a speculative Chip Test, wherein an individual's brain would be gradually replaced with durable microchips. If this individual continues to report having phenomenal consciousness, Schneider argues that this could be a reason to believe that some machines could have consciousness.

Eric Schwitzgebel, along with Mara Garza, propose four policies for the ethical design of human-grade AI. First, given substantial uncertainty about which ethical theory is correct, we should be cautious in our handling of cases regarding artificial entities where different moral theories would produce very different ethical recommendations. Second, we should avoid creating entities if it is unclear whether they deserve full human-grade rights because it is unclear whether they are conscious or to what degree. Third, AI that merits human-grade moral consideration should be able to appreciate its own value and moral status. Fourth, AI with a human-like capacity to reflect on its values should be given an opportunity to explore, discover, and possibly alter its values.

As AIs acquire greater capacities, they are likely to acquire greater moral status, raising questions about how we should treat them. In the final chapter in the book, I sketch a theory of moral status and consider what kind of moral status an AI can have. Among other things, I argue that AIs that are alive, conscious, or sentient, or that can feel pain, have desires, or have rational or moral agency, should have the same kind of moral status as entities that have the same kind of intrinsic properties, and that a sufficient condition for an AI to have human-level moral status and be a rightsholder is when it has the physical basis for moral agency. I also consider what kind of rights a rightsholding AI could have and how AIs could have moral status greater than that of humans.

The contributions in this volume represent the state-of-the-art thinking on AI and morality from some of the leading scientists and academics in the field. AI researchers and philosophers have much to learn from each other, and a main goal of this volume is to provide a forum for this collaborative dialogue and to encourage such conversations in the future. There is a pressing need for all of us to think through these issues given the rapid development of AI. The future of humanity may depend on it.[139]

Notes

1. IBM's Watson. See "IBM and 'Jeopardy!' Relive History with Encore Presentation of 'Jeopardy!': The IBM Challenge," *Jeopardy!* website, 2011, https://web.archive.org/web/20130616092431/http://www.jeopardy.com/news/watson1x7ap4.php.
2. Google DeepMind's AlphaGo. Richard Lawler, "Google DeepMind AI Wins Final Go Match for 4-1 Series Win," *Engadget*, March 14, 2016, https://www.engadget.com/2016/03/14/the-final-lee-sedol-vs-alphago-match-is-about-to-start/.
3. "AI Beats Professionals in Six-Player Poker," *Science News*, July 11, 2019, https://www.sciencedaily.com/releases/2019/07/190711141343.htm.
4. Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "a Neural Algorithm of Artistic Style," *Computing Research Repository* (2015), https://arxiv.org/pdf/1508.06576.pdf.
5. G. Clay Whittaker, "Listen to a Song Written by Artificial Intelligence, Inspired by the Beatles," *Popular Science*, September 22, 2016, http://www.popsci.com/listen-to-this-song-by-an-artificial-intelligence-and-tell-me-its-not-beatles.
6. Martín Abadi and David G. Andersen, "Learning to Protect Communications with Adversarial Neural Cryptography," *Computing Research Repository* (2016), https://arxiv.org/pdf/1610.06918.pdf.
7. Siow-Wee Chang et al., "Oral Cancer Prognosis Based on Clinicopathologic and Genomic Markers Using a Hybrid of Feature Selection and Machine Learning Methods," *BMC Bioinformatics* 14 (2013): 170. See also Eric Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (New York: Basic Books, 2019).
8. Giuseppe Carleo and Matthias Troyer, "Solving the Quantum Many-Body Problem with Artificial Neural Networks," *Science* 355, no. 6325 (2017): 602–6.
9. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352, no. 6293 (2016): 1573–76.
10. Peter Asaro, "On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-Making," *International Review of the Red Cross* 94, no. 886 (2012): 687–709.
11. Xiaolin Wu and Xi Zhang, "Automated Inference on Criminality Using Face Images," *Computing Research Repository* (2016), https://confilegal.com/wp-content/uploads/2016/11/ESTUDIO-UNIVERSIDAD-DE-JIAO-TONG-SHANGHAI.pdf.
12. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
13. See, e.g., Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016).
14. Don Lee, "Desperate for Workers, Ageing Japan Turns to Robots for Healthcare," *The Star* (Malaysia), August 10, 2019, https://www.thestar.com.my/tech/tech-news/2019/08/10/desperate-for-workers-ageing-japan-turns-to-robots-for-healthcare.

15. Jenny Kleeman, "The Race to Build the World's First Sex Robot," *Guardian*, April 27, 2017, https://www.theguardian.com/technology/2017/apr/27/race-to-build-world-first-sex-robot; Kate Devlin, *Turned On: Science, Sex and Robots* (London: Bloomsbury, 2018).
16. Heidi Vella, "From Sex Robot to Lifelong Companion: Will We Marry Robots by 2050?," *Factor*, January 30, 2017, http://factor-tech.com/feature/from-sex-robot-to-lifelong-companion-will-we-marry-robots-by-2050/.
17. "When the Robots Take Over, Will There Be Jobs Left for Us?," *CBS Sunday Morning*, April 9, 2017, http://www.cbsnews.com/news/when-the-robots-take-over-will-there-be-jobs-left-for-us/.
18. Kevin J. Delaney, "The Robot That Takes Your Job Should Pay Taxes, Says Bill Gates," *Quartz*, February 17, 2017, https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/.
19. David Brancaccio, "What Universal Basic Income Could Mean for the Future of Work," *Marketplace*, April 18, 2017, https://www.marketplace.org/2017/04/18/economy/robot-proof-jobs/basic-income-y-combinator-oakland-krisiloff.
20. Ray Kurzweil, *The Singularity Is Near* (New York: Viking Press, 2005).
21. Alan Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–60. John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–24.
22. Searle, "Minds, Brains, and Programs."
23. Vernor Vinge, "The Coming Technological Singularity," *Whole Earth Review*, Winter (1993), https://edoras.sdsu.edu/~vinge/misc/singularity.html; David Chalmers, "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies* 17, nos. 9–10 (2010): 7–65.
24. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
25. Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan Ćirković, 91-119 (Oxford: Oxford University Press, 2008).
26. S. Matthew Liao, "The Basis of Human Moral Status," *Journal of Moral Philosophy* 7 (2010): 159–79.
27. Eric Schwitzgebel and Mara Garza, "A Defense of the Rights of Artificial Intelligences," *Midwest Studies in Philosophy* 39, no. 1 (2015): 98–119.
28. See, e.g., Bostrom, *Superintelligence*; Jerry Kaplan, *Artificial Intelligence: What Everyone Needs to Know* (New York: Oxford University Press, 2016)., Patrick Lin, Keith Abney, and Ryan Jenkins, eds., *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (New York: Oxford University Press, 2017); Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).
29. This section is intended to be accessible to a nontechnical audience.
30. John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

31. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice Hall, 2010), 2.
32. The discussion below draws on ibid., which provides a good overview of different types of AI algorithms. For another perspective, see Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (New York: Basic Books, 2015).
33. John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, MA: MIT Press, 1985).
34. Russell and Norvig, *Artificial Intelligence*, 696.
35. Some classification algorithms are support vector machines (which output an optimal hyperplane that categorizes new examples), naive Bayes (which assume that every feature is independent), and nearest neighbors (which try to find the point in a given set that is closest, or most similar, to a given point).
36. Some regression algorithms are linear regression (which models the relationship between dependent and independent variables), support vector regression (which applies the support vector method to regression problems), and regression trees (which are decisions trees where the target variable takes continuous values).
37. Some clustering algorithms include k-means clustering (where data points are clustered into a number of mutually exclusive clusters) and hierarchical clustering (where some kind of hierarchy is imposed on data points).
38. Russell and Norvig, *Artificial Intelligence*, ch. 21.
39. Frank Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* 65, no. 6 (1958): 386.
40. See, e.g., D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, ed. E. Rumelhart David, L. McClelland James, and Corporate Pdp Research Group (Cambridge, MA: MIT Press, 1986), 318–62.
41. See, e.g., Yann LeCun, "Une Procedure D'apprentissage Pour Reseau a Seuil Asymmetrique (A Learning Scheme for Asymmetric Threshold Networks)," *Proceedings of Cognitiva* 85 (1985): 599–604
42. See, e.g., Leon Bottou et al., "High Quality Document Image Compression with 'Djvu,'" *Journal of Electronic Imaging* 7, no. 3 (1998): 410–25.
43. For an account of the development of deep learning, see Jürgen Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks* 61 (2015): 85–117.
44. Paul Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences" (PhD diss., Harvard University, 1974).
45. David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors," *Nature* 323, no. 6088 (1986): 533–36.
46. Y. LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation* 1, no. 4 (1989): 541–51.
47. David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016): 484.

48. Kevin Granville, "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens," *New York Times*, March 19, 2018, https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.
49. Megan Moteni, "23andMe's Pharma Deals Have Been the Plan All Along," *Wired*, August 3, 2018, https://www.wired.com/story/23andme-glaxosmithkline-pharma-deal/.
50. Tom Simonite, "When It Comes to Gorillas, Google Photos Remain, Blind," *Wired*, January 11, 2018, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.
51. Wendy Lee, "How Tech's Lack of Diversity Leads to Racist Software," SFGate, updated July 22, 2015, https://www.sfgate.com/business/article/How-tech-s-lack-of-diversity-leads-to-racist-6398224.php.
52. "Model Selection, Underfitting and Overfitting," in *Dive into Deep Learning*, accessed February 20, 2020, https://www.d2l.ai/chapter_multilayer-perceptrons/underfit-overfit.html.
53. Ibid.
54. Colin Lecher, "What Happens When an Algorithm Cuts Your Health Care," *Verge*, March 21, 2018, https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.
55. Zachary C. Lipton, "The Mythos of Model Interpretability," *Queue* 16, no. 3 (2018): 31–57.
56. See, e.g., Christoph Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," December 17, 2019, https://christophm.github.io/interpretable-ml-book/.
57. See, e.g., ibid.
58. Some such techniques include permutation importance, which considers when a model produces counterintuitive results, and partial dependence plot, which considers the marginal effect that one or two features have on a prediction. Ibid.
59. For an argument that post hoc reasoning need not always be biased, see S. Matthew Liao, "Morality and Neuroscience: Past and Future," in *Moral Brains: The Neuroscience of Morality*, ed. S. Matthew Liao (New York: Oxford University Press, 2016), 1–42.
60. Lipton, "The Mythos of Model Interpretability"; Alex John London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *Hastings Center Report* 49, no. 1 (2019): 15–21.
61. See, e.g., London, "Artificial Intelligence and Black-Box Medical Decisions," who expresses this concern.
62. Ibid.
63. Mohd Shara and Sidney J. Stohs, "Efficacy and Safety of White Willow Bark (Salix Alba) Extracts," *Phytotherapy Research* 29, no. 8 (2015): 1112–16.
64. K. L. Kopnisky et al., "Chronic Lithium Treatment Antagonizes Glutamate-Induced Decrease of Phosphorylated CREB in Neurons via Reducing Protein Phosphatase 1 and Increasing MEK Activities," *Neuroscience* 116, no. 2 (2003): 425–35.

65. For a discussion concerning the importance of developing machine learning models that can capture causal relations, see, e.g., Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York: Basic Books, 2018).
66. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv*, 2014, https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6572G; N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access* 6 (2018): 14410–30.
67. Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One Pixel Attack for Fooling Deep Neural Networks," *arXiv*, 2017, https://ui.adsabs.harvard.edu/abs/2017arXiv171008864S.
68. Ibid.
69. Goodfellow, Shlens, and Szegedy, "Explaining and Harnessing Adversarial Examples."
70. Dan Hendrycks et al., "Natural Adversarial Examples," *arXiv*, 2019, https://ui.adsabs.harvard.edu/abs/2019arXiv190707174H.
71. Searle, "Minds, Brains, and Programs."
72. Wendell Wallach and Colin Allen, eds., *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2008); Michael Anderson and Susan Leigh Anderson, eds., *Machine Ethics* (Cambridge: Cambridge University Press, 2011).
73. Thomas Powers, "Prospects for a Kantian Machine," *Intelligent Systems, IEEE* 21 (2006): 46–51.
74. See, e.g., Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York: Oxford University Press, 2016).
75. Philippa Foot, "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review* 5 (1967): 5–15.
76. Judith Jarvis Thomson, "The Trolley Problem," *Yale Law Journal* 94 (1985): 1395–415. The term "Trolley Problem" is sometimes used in the machine ethics literature to refer to any case that resembles Sidetrack and Footbridge. Strictly speaking, however, many of the cases discussed in the machine ethics literature are structurally different from Sidetrack and Footbridge. See Kamm, in this volume, for a more in-depth discussion of this point.
77. See Frances Myra Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007), 93, for this formulation.
78. See, e.g., Thomson, "The Trolley Problem."
79. Kamm, *Intricate Ethics*, defends what might be called a causal structure theory.
80. Edmond Awad et al., "The Moral Machine Experiment," *Nature* 563, no. 7729 (2018): 59–64.
81. Joshua D. Greene, "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics," in *Moral Brains: The Neuroscience of Morality*, ed. S. Matthew Liao (New York: Oxford University Press, 2016), 119–49.
82. Lindsay Dodgson, "Why Mercedes Plans to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations," *Business Insider*, October 12, 2016, https://www.businessinsider.com/mercedes-benz-self-driving-cars-programmed-save-driver-2016-10.

83. Apparently, China is already building such highways. Dan Robitzski, "China's Rolling Out Dedicated Highway Lanes for Self-Driving Cars," *Futurism*, April 23, 2019, https://futurism.com/the-byte/china-dedicated-highway-lanes-self-driving-cars.
84. Yujing Liu, "Facial Recognition Tech Catches Fugitive in Huge Crowd at Jacky Cheung Cantopop Concert in China," *South China Morning Post*, April 12, 2018, https://www.scmp.com/news/china/society/article/2141387/facial-recognition-tech-catches-fugitive-among-huge-crowd-pop.
85. Michal Kosinski and Yilun Wang, "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images," *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246–57. This study is controversial because it can be questioned whether one can determine sexual orientation by photos.
86. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," paper presented at the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, November 27–30, 2018.
87. Aja Romano, "Jordan Peele's Simulated Obama PSA Is a Double-Edged Warning against Fake News," *Vox*, April 18, 2018, https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed.
88. Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War," *Foreign Affairs*, January–February 2019, https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war?cid=otr-authors-january_february_2019-121118.
89. Jack Morse, "Fake 'Drunk' Nancy Pelosi Video Goes Viral, and It Wasn't Even That Hard to Make," *Mashable*, May 23, 2019, https://mashable.com/article/nancy-pelosi-edited-video-sound-drunk-deepfakes/.
90. James Manyika et al., *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation* (San Francisco: McKinsey Global Institute, 2017).
91. Carl Benedikt Frey and Michael A. Osborne, "The Future of Employment: How Susceptible Are Jobs to Computerisation?," *Technological Forecasting and Social Change* 114 (2017): 254–80.
92. Till Alexander Leopold, Saadia Zahid, and Vesselina Ratcheva, *The Future of Jobs Report* (New York: World Economic Forum, 2018).
93. Erik Brynjolfsson, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," *AEA Papers and Proceedings* 108 (2018): 43–47.
94. *BBC News*, https://www.bbc.com/news/stories-49343280
95. David Levy, *Love and Sex with Robots: The Evolution of Human-Robot Relationships* (New York: HarperCollins, 2007).
96. https://www.care-o-bot-4.de/
97. Adam Satariano, Elian Peltier, and Dmitry Kostyukov, "Meet Zora, the Robot Caregiver," *New York Times*, November 23, 2018, https://www.nytimes.com/interactive/2018/11/23/technology/robot-nurse-zora.html.
98. A. John Simmons, "Human Rights and World Citizenship: The Universality of Human Rights in Kant and Locke," in *Justification and Legitimacy: Essays on Rights and Obligations* (Cambridge: Cambridge University Press, 2001), 185.

99. This discussion draws on S. Matthew Liao, "Human Rights as Fundamental Conditions for a Good Life," in *Philosophical Foundations of Human Rights*, ed. Rowan Cruft, S. Matthew Liao, and Massimo Renzo (Oxford: Oxford University Press, 2015), 79–100.
100. Maurice Cranston, *What Are Human Rights?* (London: Bodley Head, 1973).
101. Scott Shane and Daisuke Wakabayashi, "'The Business of War': Google Employees Protest Work for the Pentagon," *New York Times*, April 4, 2018, https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.
102. Daisuke Wakabayashi and Scott Shane, "Google Will Not Renew Pentagon Contract That Upset Employees," *New York Times*, June 1, 2018, https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.
103. Massimo Calabresi, "Inside Russia's Social Media War on America," *Time*, updated May 18, 2017, https://time.com/4783932/inside-russia-social-media-war-america/.
104. Farhad Manjoo, "A Plan in Case Robots Take the Jobs: Give Everyone a Paycheck," *New York Times*, March 2, 2016, https://www.nytimes.com/2016/03/03/technology/plan-to-fight-robot-invasion-at-work-give-everyone-a-paycheck.html.
105. Sigal Samuel, "Finland Gave People Free Money: It Increased Their Trust in Social Institutions," *Vox*, April 6, 2019, https://www.vox.com/2019/4/6/18297452/finland-basic-income-free-money-canada.
106. "About the Administration on Aging (AoA)," Administration for Community Living, accessed February 20, 2020, http://aoa.acl.gov/Aging_Statistics/Profile/2014/docs/2014-Profile.pdf.
107. Eric Klinenberg, Stacy Torres, and Elena Portacolone, "Aging Alone in America," briefing paper prepared for the Council on Contemporary Families for Older Americans Month, May 1, 2013, http://www.contemporaryfamilies.org/wp-content/uploads/2013/10/2012_Briefing_Klinenberg_Aging-alone-in-america.pdf.
108. Laura Fratiglioni et al., "Influence of Social Network on Occurrence of Dementia: A Community-Based Longitudinal Study," *Lancet* 355, no. 9212 (2000): 1315–19.
109. L. L. Barnes et al., "Social Resources and Cognitive Decline in a Population of Older African Americans and Whites," *Neurology* 63, no. 12 (2004): 2322–26.
110. Future of Life Institute, "Asilomar AI Principles," accessed February 20, 2020, https://futureoflife.org/ai-principles/.
111. Partnership on AI, "Tenets," accessed February 20, 2020, https://www.partnershiponai.org/tenets/.
112. European Commission, "Ethics Guidelines for Trustworthy Artificial Intelligence," June 2018, https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.
113. Bostrom, *Superintelligence*. It is sometimes also called the AI Safety problem. For a good collection addressing this problem, see Roman V. Yampolskiy, ed., *Artificial Intelligence Safety and Security* (London: Chapman & Hall, 2018).
114. Pathmind, *A.I. Wiki*, accessed February 20, 2020, https://skymind.ai/wiki/strong-ai-general-ai.

115. There is also "human-level" *narrow* AI, which would involve an AI's being able to perform at a specific task at the human level. Arguably, programs such as AlphaGo and IBM Watson are human-level narrow AIs. They might even qualify as super human narrow AIs. I thank David Chalmers for this point.
116. Nick Bostrom, for instance, takes an artificial superintelligence to be "any intellect that greatly exceeds the *cognitive performance* of humans in virtually all domains of interest" (*Superintelligence*, 22, italics added). This could be interpreted as taking a superintelligent machine to have cognitive but not emotional or moral intelligence.
117. Searle, "Minds, Brains, and Programs"; Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992).
118. Luciano Floridi, "Should We Be Afraid of AI?," *Aeon*, May 9, 2016, https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible.
119. Russell and Norvig, *Artificial Intelligence*.
120. See, e.g., Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).
121. Thomas Nagel, "What Is It Like to Be a Bat?," *Philosophical Review* 83, no. 4 (1974): 435–50.
122. S. Matthew Liao, "Twinning, Inorganic Replacement, and the Organism View," *Ratio* 23, no. 1 (2010): 59–72.
123. For a discussion of these three possibilities, see Chalmers, "The Singularity."
124. See, e.g., Isaac Asimov, "Runaround," in *I, Robot*, 1942).
125. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk"; Russell and Norvig, *Artificial Intelligence*.
126. For instance, these laws fail properly to define and distinguish "humans" and "robots." Also, robots could unknowingly breach the laws if they did not have full information.
127. Several contributions in part III of this volume are devoted to this problem.
128. Bostrom, *Superintelligence*, offers a good overview of this and other methods of controlling a superintelligent AI.
129. Bostrom refers to a version of this method as the tripwire method (ibid.).
130. M. Kosfeld et al., "Oxytocin Increases Trust in Humans," *Nature* 435, no. 7042 (2005): 673–76.
131. See, e.g., Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra, "Brain-Computer Interface Systems and Their Applications," in *Introduction to EEG- and Speech-Based Emotion Recognition*, ed. Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra (London: Academic Press, 2016).
132. Gabriella Deli et al., "Deep Brain Stimulation Can Preserve Working Status in Parkinson's Disease," *Parkinson's Disease* 2015, article 936965, https://www.hindawi.com/journals/pd/2015/936865/.
133. Steffen Steinert et al., "Doing Things with Thoughts: Brain-Computer Interfaces and Disembodied Agency," *Philosophy & Technology* (2018): 457-482.
134. John F. Burke et al., "Brain Computer Interface to Enhance Episodic Memory in Human Participants," *Frontiers in Human Neuroscience* 8, no. 1055 (2015): 1–10.

135. André Russowsky Brunoni, and Marie-Anne Vanderhasselt, "Working Memory Improvement with Non-Invasive Brain Stimulation of the Dorsolateral Prefrontal Cortex: A Systematic Review and Meta-Analysis," *Brain and Cognition* 86 (2014): 1–9; Anke Hammer et al., "Errorless and Errorful Learning Modulated by Transcranial Direct Current Stimulation," *BMC Neuroscience* 12, no. 1 (2011): 72; Djamila Bennabi et al., "Transcranial Direct Current Stimulation for Memory Enhancement: From Clinical Research to Animal Models," *Frontiers in Systems Neuroscience* 8, no. 159 (2014): 1-9.
136. Clement Hamani et al., "Memory Enhancement Induced by Hypothalamic/Fornix Deep Brain Stimulation," *Annals of Neurology* 63, no. 1 (2008): 119–23; Nanthia Suthana and Itzhak Fried. "Deep Brain Stimulation for Enhancement of Learning and Memory," *NeuroImage* 85 (2014): 996–1002.
137. Research of such closed-loop systems is already underway. See "Wireless 'Pacemaker for the Brain' Could Offer New Treatment for Neurological Disorders," January 1, 2019, *Science Daily*, https://www.sciencedaily.com/releases/2019/01/190101094517.htm. Also, DARPA, the Defense Advanced Research Projects Agency, has a program called Systems-Based Neurotechnology for Emerging Therapies, the mission of which is to create the "next-generation, closed-loop neural stimulators that exceed currently developed capacities for simultaneous stimulation and recording, with the goal of providing investigators and clinicians an unprecedented ability to record, analyze, and stimulate multiple brain regions for therapeutic purposes." Al Emondi, "Systems-Based Neurotechnology for Emerging Therapies (SUBNETS)," DARPA, accessed February 20, 2020, https://www.darpa.mil/program/systems-based-neurotechnology-for-emerging-therapies.
138. John Markoff, "Elon Musk's Neuralink Wants 'Sewing Machine–Like' Robots to Wire Brains to the Internet," *New York Times*, July 16, 2019, https://www.nytimes.com/2019/07/16/technology/neuralink-elon-musk.html.
139. I would like to thank David Chalmers, Wibke Gruetjen, Nicola Allais, Eddy Nahmias, Adam Birt, Nicholas Tilmes, Carissa Véliz, and Kimi Chernoby for their very helpful comments on earlier versions of this piece.

References

Abadi, Martín, and David G. Andersen. "Learning to Protect Communications with Adversarial Neural Cryptography." *Computing Research Repository* (2016). https://arxiv.org/pdf/1610.06918.pdf.

Abhang, Priyanka A., Bharti W. Gawali, and Suresh C. Mehrotra. "Brain-Computer Interface Systems and Their Applications." In *Introduction to EEG- and Speech-Based Emotion Recognition*, edited by Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra, 165–77. London: Academic Press, 2016.

Akhtar, N., and A. Mian. "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey." *IEEE Access* 6 (2018): 14410–30.

Anderson, Michael, and Susan Leigh Anderson, eds. *Machine Ethics*. Cambridge: Cambridge University Press, 2011.

Asaro, Peter. "On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-Making." *International Review of the Red Cross* 94, no. 886 (2012): 687–709.

Asimov, Isaac. "Runaround." In *I, Robot*. New York: Bantam Dell, 1942.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563, no. 7729 (2018): 59–64.

Barnes, L. L., C. F. Mendes de Leon, R. S. Wilson, J. L. Bienias, and D. A. Evans. "Social Resources and Cognitive Decline in a Population of Older African Americans and Whites." *Neurology* 63, no. 12 (2004): 2322–26.

Bennabi, Djamila, Solène Pedron, Emmanuel Haffen, Julie Monnin, Yvan Peterschmitt, and Vincent Van Waes. "Transcranial Direct Current Stimulation for Memory Enhancement: From Clinical Research to Animal Models." *Frontiers in Systems Neuroscience* 8, no. 159 (2014): 1-9.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. "The Social Dilemma of Autonomous Vehicles." *Science* 352, no. 6293 (2016): 1573–76.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Bottou, Leon, Patrick Haffner, Paul G. Howard, Patrice Simard, Yoshua Bengio, and Yann LeCun1. "High Quality Document Image Compression with 'Djvu.'" *Journal of Electronic Imaging* 7, no. 3 (1998): 410–25.

Brunoni, André Russowsky, and Marie-Anne Vanderhasselt. "Working Memory Improvement with Non-Invasive Brain Stimulation of the Dorsolateral Prefrontal Cortex: A Systematic Review and Meta-Analysis." *Brain and Cognition* 86 (2014): 1–9.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108 (2018): 43–47.

Burke, John F., Maxwell B. Merkow, Joshua Jacobs, Michael J. Kahana, and Kareem A. Zaghloul. "Brain Computer Interface to Enhance Episodic Memory in Human Participants." *Frontiers in Human Neuroscience* 8, no. 1055 (2015): 1-10.

Carleo, Giuseppe, and Matthias Troyer. "Solving the Quantum Many-Body Problem with Artificial Neural Networks." *Science* 355, no. 6325 (2017): 602–6.

Chalmers, David. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, nos. 9–10 (2010): 7–65.

Chang, Siow-Wee, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain. "Oral Cancer Prognosis Based on Clinicopathologic and Genomic Markers Using a Hybrid of Feature Selection and Machine Learning Methods." *BMC Bioinformatics* 14 (2013): 170.

Cranston, Maurice. *What Are Human Rights?* London: Bodley Head, 1973.

Deli, Gabriella, István Balás, Tamás Dóczi, József Janszky, Kázmér Karádi, Zsuzsanna Aschermann, Ferenc Nagy, Attila Makkos, Márton Kovács, Edit Bosnyák, Norbert Kovács, and Sámuel Komoly. "Deep Brain Stimulation Can Preserve Working Status in Parkinson's Disease." *Parkinson's Disease* 2015, article 936965. https://www.hindawi.com/journals/pd/2015/936865/.

Devlin, Kate. *Turned On: Science, Sex and Robots*. London: Bloomsbury, 2018.
Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, 2015.
Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1992.
Foot, Philippa. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5 (1967): 5–15.
Fratiglioni, Laura, Hui-Xin Wang, Kjerstin Ericsson, Margaret Maytan, and Bengt Winblad. "Influence of Social Network on Occurrence of Dementia: A Community-Based Longitudinal Study." *Lancet* 355, no. 9212 (2000): 1315–19.
Frey, Carl Benedikt, and Michael A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114 (2017): 254–80.
Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A Neural Algorithm of Artistic Style." *Computing Research Repository* (2015), https://arxiv.org/pdf/1508.06576.pdf
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." *arXiv*, 2014. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6572G.
Greene, Joshua D. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics." In *Moral Brains: The Neuroscience of Morality*, edited by S. Matthew Liao, 119–49. New York: Oxford University Press, 2016.
Güera, D., and E. J. Delp. "Deepfake Video Detection Using Recurrent Neural Networks." Paper presented at the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, November 27–30, 2018.
Hamani, Clement, Mary Pat McAndrews, Melanie Cohn, Michael Oh, Dominik Zumsteg, Colin M. Shapiro, Richard A. Wennberg, and Andres M. Lozano. "Memory Enhancement Induced by Hypothalamic/Fornix Deep Brain Stimulation." *Annals of Neurology* 63, no. 1 (2008): 119–23.
Hammer, Anke, Bahram Mohammadi, Marlen Schmicker, Sina Saliger, and Thomas F. Münte. "Errorless and Errorful Learning Modulated by Transcranial Direct Current Stimulation." *BMC Neuroscience* 12, no. 1 (2011): 72.
Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press, 1985.
Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. "Natural Adversarial Examples." *arXiv*, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190707174H.
Kamm, Frances Myra. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.
Kaplan, Jerry. *Artificial Intelligence: What Everyone Needs to Know*. New York: Oxford University Press, 2016.
Kopnisky, K. L., E. Chalecka-Franaszek, M. Gonzalez-Zulueta, and D. M. Chuang. "Chronic Lithium Treatment Antagonizes Glutamate-Induced Decrease of Phosphorylated Creb in Neurons via Reducing Protein Phosphatase 1 and Increasing Mek Activities." *Neuroscience* 116, no. 2 (2003): 425–35.
Korsgaard, Christine M. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.
Kosfeld, M., M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr. "Oxytocin Increases Trust in Humans." *Nature* 435, no. 7042 (2005): 673–76.

Kosinski, Michal, and Yilun Wang. "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images." *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246–57.

Kurzweil, Ray. *The Singularity Is Near*. New York: Viking Press, 2005.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1, no. 4 (1989): 541–51.

LeCun, Yann. "Une Procedure D'apprentissage Pour Reseau a Seuil Asymmetrique (A Learning Scheme for Asymmetric Threshold Networks)." *Proceedings of Cognitiva* 85 (1985): 599–604.

Leopold, Till Alexander, Saadia Zahid, and Vesselina Ratcheva. *The Future of Jobs Report*. New York: World Economic Forum, 2018.

Levy, David. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: HarperCollins, 2007.

Liao, S. Matthew. "The Basis of Human Moral Status." *Journal of Moral Philosophy* 7 (2010): 159–79.

Liao, S. Matthew. "Human Rights as Fundamental Conditions for a Good Life." In *Philosophical Foundations of Human Rights*, edited by Rowan Cruft, S. Matthew Liao, and Massimo Renzo, 79–100. Oxford: Oxford University Press, 2015.

Liao, S. Matthew. "Morality and Neuroscience: Past and Future." In *Moral Brains: The Neuroscience of Morality*, edited by S. Matthew Liao, 1–42. New York: Oxford University Press, 2016.

Liao, S. Matthew. "Twinning, Inorganic Replacement, and the Organism View." *Ratio* 23, no. 1 (2010): 59–72.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, 2017.

Lipton, Zachary C. "The Mythos of Model Interpretability." *Queue* 16, no. 3 (2018): 31–57.

London, Alex John. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49, no. 1 (2019): 15–21.

Manyika, James, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi. *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation.* San Francisco: McKinsey Global Institute, 2017.

McCarthy, John, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." August 31, 1955. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

Molnar, Christoph. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." December 17, 2019. https://christophm.github.io/interpretable-ml-book/.

Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4 (1974): 435–50.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*: New York: Crown, 2016.

Pearl, Judea, and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books, 2018.

Powers, Thomas. "Prospects for a Kantian Machine." *Intelligent Systems, IEEE* 21 (2006): 46–51.

Rosenblatt, Frank. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65, no. 6 (1958): 386.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, edited by E. Rumelhart David, L. McClelland James, and Corporate Pdp Research Group, 318–62. Cambridge, MA: MIT Press, 1986.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-Propagating Errors." *Nature* 323, no. 6088 (1986): 533–36.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

Schmidhuber, Jürgen. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61 (2015): 85–117.

Schwitzgebel, Eric, and Mara Garza. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39, no. 1 (2015): 98–119.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–24.

Shara, Mohd, and Sidney J. Stohs. "Efficacy and Safety of White Willow Bark (Salix Alba) Extracts." *Phytotherapy Research* 29, no. 8 (2015): 1112–16.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (2016): 484.

Simmons, A. John. "Human Rights and World Citizenship: The Universality of Human Rights in Kant and Locke." In *Justification and Legitimacy: Essays on Rights and Obligations*, 179–96. Cambridge: Cambridge University Press, 2001.

Steinert, Steffen, Christoph Bublitz, Ralf Jox, and Orsolya Friedrich. "Doing Things with Thoughts: Brain-Computer Interfaces and Disembodied Agency." *Philosophy & Technology* 32 (2018): 457-482.

Su, Jiawei, Danilo Vasconcellos Vargas, and Sakurai Kouichi. "One Pixel Attack for Fooling Deep Neural Networks." *arXiv*, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv171008864S.

Suthana, Nanthia, and Itzhak Fried. "Deep Brain Stimulation for Enhancement of Learning and Memory." *NeuroImage* 85 (2014): 996–1002.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Thomson, Judith Jarvis. "The Trolley Problem." *Yale Law Journal* 94 (1985): 1395–415.

Topol, Eric. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books, 2019.

Turing, Alan. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–60.

Vallor, Shannon. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press, 2016.

Vinge, Vernor. "The Coming Technological Singularity." *Whole Earth Review*, Winter (1993), https://edoras.sdsu.edu/~vinge/misc/singularity.html.

Wallach, Wendell, and Colin Allen, eds. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.

Werbos, Paul. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral *Sciences*." PhD diss., Harvard University, 1974.

Wu, Xiaolin, and Xi Zhang. "Automated Inference on Criminality Using Face Images." *Computing Research Repository* (2016), https://confilegal.com/wp-content/uploads/2016/11/ESTUDIO-UNIVERSIDAD-DE-JIAO-TONG-SHANGHAI.pdf.

Yampolskiy, Roman V., ed. *Artificial Intelligence Safety and Security*. London: Chapman & Hall, 2018.

Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Ćirković, 91-119. Oxfords: Oxford University Press, 2008.

Part I: Building Ethics Into Machines

1
Ethical Learning, Natural and Artificial

Peter Railton

1.1. Introduction

There is no shortage of urgent ethical questions about the responsible development and deployment of artificial intelligence. Artificial intelligence is a *fundamental* technological innovation in the sense that, besides adding new technological possibilities of its own, it alters the capabilities and potential benefits and risks of a wide range of other technologies, including such "soft technologies" as social practices and institutions. One has only to imagine an area of human life—work, communication, governance, mobility, medicine, warfare—in order to have exciting *and* disturbing potential effects of artificial intelligence spring to mind, which grow more exciting and more disturbing the further one imagines artificial intelligence to be capable of developing.

This paper will not address directly any of these particular problems about the responsible development and deployment of artificial intelligence, important as they are; rather it considers a question that could be relevant to all of them, since it concerns the nature of artificial intelligence itself. Increasingly, artificial systems will be exercising life-affecting functions or making life-affecting decisions—in piloting a vehicle, in home healthcare, in hiring and firing, in monitoring and shaping the information we receive—that we would not ordinarily entrust to someone lacking in sensitivity to ethical concerns. How, then, might artificial systems come to be appropriately sensitive to ethical concerns? Moreover, how might such sensitivity be a core part of their intelligence and capacities? My primary focus will be on how this might be possible. To a first approximation, we can characterize sensitivity to ethical concerns as a robust, reliable capacity to detect and respond appropriately to ethically relevant features of situations, actions, agents, and outcomes. Our answer to questions about how we can responsibly develop or deploy artificial systems will depend significantly upon the extent to which such apt responsiveness to ethically relevant features is possible.

A closely related question, to my mind, is this: As artificial intelligence becomes more general and capable, it will give rise not only to new technological possibilities, but to new classes of *agents* that operate independently of direct

Peter Railton, *Ethical Learning, Natural and Artificial* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905033.003.0002

human supervision or control, and with which we can increasingly have *social* rather than merely instrumental relations. For example, as a matter of safety, it may be important for artificial agents to be able to refuse to comply with certain commands, or to have an element of uncertainty about the goals we give them, so that they pay attention to accumulating evidence of harms, bias, or dysfunction and can make their own decision to suspend pursuit of such goals or seek more information and advice. While there are dangers inherent in creating highly capable artificial agents with enough autonomy to question the goals they are given on grounds of harm, bias, or dysfunction, there is greater danger in creating highly capable artificial agents lacking any capacity to do so. Think only of the same issue raised with respect to raising human (and presumably highly capable) agents; as we will see, human infant ethical development typically proceeds in a sufficiently autonomous way that three- to four-year-olds will question rules given to them by persons in authority when they believe these rules to cause inappropriate harm or be unfair.[1] Moreover, often the information artificial agents need to make such decisions will be available most rapidly or reliably from other artificial agents. We will need to find ways to coordinate, cooperate, collaborate, and compete peacefully and productively with artificial systems, seen as independent parties whose behavior we cannot simply dictate. Moreover, artificial systems will need to find ways to coordinate, cooperate, collaborate, and compete peacefully and productively with us in return, and with each other. Artificial systems capable of projecting and evaluating future courses of action, of assessing benefits and harms to self and other, of making commitments, and of regulating their own behavior accordingly will be capable of something like social-contract reasoning: we could negotiate with them terms of mutually beneficial cooperation that all of us would constrain ourselves to follow.

This capacity for social-contract reasoning, and for mutual constraint for mutual benefit, does not presuppose a capacity for qualitative experiences or emotions akin to humans. In humans, our actively norm-governed life together is greatly enhanced by our capacity for a range of affective states—empathic simulation and emotions such as loyalty, guilt, forgiveness, and so on[2]—but an actively norm-governed life does not seem to require such feelings, so long as there are sufficiently developed agential capacities for self-regulation, representation of others' goals and information, and the formation of conventions, agreements, or commitments. Even before we have to contend with possible "super-intelligences,"[3] we will need to ask how to contend with artificial agents from whose capacities we could greatly benefit but whose cooperation with us will not be entirely up to us and may depend upon negotiation in which we seek to find common ground for working together and according to each other's goals. Indeed, for intelligent systems to be able to robustly and reliably detect and respond to ethically relevant features they may need to have at least this much

autonomy in deciding whether to work with the particular human or artificial agents who might seek to control them, and for what purposes.

We are not without experience of highly capable nonhuman agents lacking a unified consciousness or affective states but possessing extraordinary levels of information and problem-solving ability, and whose aims may differ from our own in ways that require us to negotiate with them if we are to gain the benefits they make possible. Corporate entities—governments, corporations, universities, institutes, unions, political parties—can have a distinctive set of goals related to their own purposes or conditions for survival and flourishing, which may overlap with but also fail to be the same as those of the individuals who compose them or are affected by them. They possess capacities for pursuing values and holding themselves to norms, for future projection and planning, for entering into (or failing to enter into) cooperative arrangements, strategic alliances, mutual commitments or contracts, and for incurring, carrying out, and policing compliance with associated obligations. At the same time, they are not fully transparent in their inner processes; asking how an action by a corporate entity came to be taken may not yield a determinate decision-process with clear lines of responsibility. Asking how we might enter into mutually beneficial, mutually constrained, normatively governed relations with emerging agents possessing higher-than-human intelligence is like asking how we are able to enter into such relations with governments, corporations, and so on. We have made a fair amount of progress in developing countervailing institutions and normative practices that enable us to work with such agents in ways that can be mutually beneficial. But this is still a work in progress.

1.2. A Social Perspective

Artificial neural networks, I am told, were originally inspired by the thought that naturally occurring cortical architecture is the result of countless generations of selection for a capacity to learn and act intelligently, and so is a plausible basis upon which to build artificial intelligence. Now that artificial neural networks have become sufficiently deep and fast, and data have become sufficiently plentiful, this inspiration is bearing fruit. Perhaps, as we look forward to the development of *general* artificial intelligence, we should look for inspiration at the most distinctive characteristics of the naturally occurring creatures that appear to have achieved the highest levels of general intelligence: humans. And humankind is at least as distinctive, relative to the wider animal world, for our *social* capacities as for our sheer intellect. Many animals, of course, live in complex social groups, but humans are unusual in the extent of their capacity for large-scale coordination and cooperation with nonkin, open-ended exchange of information,

and normative self-regulation in light of long-term, abstract impersonal goals. If one thinks of general intelligence as a capacity for open-ended problem-solving, then our capacities for building and sustaining shared practices to solve shared problems are central to our general intelligence. These capacities enable us to create *epistemic* as well as ethical communities, leveraging our individual abilities in ways that can carry us far beyond anything we could accomplish as individual agents or inquirers.

This paper is an exploration of the idea that the project of building highly effective, generally intelligent *artificial* epistemic agents should be seen as connected with building artificial agents capable of apt responsiveness to ethically relevant features. This idea has a certain advantage in thinking about ethics and artificial intelligence. If we imagine that achieving responsiveness to ethically relevant features in an artificial system is a matter of *adding* a novel capacity or set of principles to an already fully formed general intelligence, then it might also be imagined that this capacity or these principles could readily be *subtracted* from such an intelligence without cognitive loss. If, instead, there is a root connection between full development of the capacity to be appropriately responsive to epistemically relevant features and full development of the capacity to be appropriately responsive to ethically relevant features, then responsiveness to ethically relevant features could be a *deep* feature of artificial systems with high general intelligence and problem-solving ability, not easily removed without serious impairment of other aspects of general intelligence and problem-solving. Suggestive evidence comes from the way in which some psychological disorders that seem to impair the development of appropriate responsiveness to ethically relevant features, such as psychopathy, tend also to have costs to the full development of more general human intelligence, understood as problem-solving ability.[4] Contrary to the popular idea of the psychopath as the height of rationality, intelligence, control, and savvy is the research indicating that psychopaths show serious deficits in attention, impulse-control, and ability to accurately represent likely negative future outcomes—for themselves as well as others. These deficits then help explain the difficulty of psychopaths in holding themselves to long-term goals, plans, or relationships.[5] But the point is not solely about individuals. Persons with "Machiavellian" personal disorders may be quite intelligent by conventional measures and can achieve considerable long-term success by taking advantage of the vulnerabilities of ordinary agents and practices,[6] but it is one thing to be able to exploit an existing epistemic community and another thing to be able to build and sustain one that is as effective as possible at gaining knowledge. The latter is closer to the ultimate goal of artificial intelligence research. For example, individuals scoring high in Machiavellian psychological profiles are more likely to adopt an "economically rational" strategy in trust games, whereas no known human community has this as the predominant

disposition,[7] and the communities, small and large, that most effectively work together to enlarge their capacities depart the most from this strategy.[8] Imagine the difference in learning to drive safely in an open-ended array of situations if a community of self-driving cars shares individual driving data rather than each using the data it acquires to gain whatever strategic advantages it can over the others.

It is important to distinguish our question about responsiveness to ethically relevant features from asking how artificial systems might come to have the distinctive qualitative experiences or affective responses of typical human moral agents. For example, empathic simulation appears to play an important role in helping humans to understand one another,[9] but an artificial system can engage in empathic simulation without "reliving the experience" of others if artificial systems can become sufficiently skilled at modeling others' internal states on the basis of observed behavior and can accord intrinsic weight to others' imputed goals or utility functions in evaluating simulated courses of action during decision-making.

Consider by analogy the fact that developing artificial systems capable of being effectively responsive to a range of semantically relevant features in natural language need not await the development in such systems of the full range of human thought and feeling. We might think that no artificial system could grasp the full meaning of "Where are the snows of yesteryear?" without feeling a pang of nostalgia, but a genuinely intelligent artificial system might nonetheless be able to represent the essential semantic features of this sentence and to capture enough of the pragmatics of English usage to give it as a suitable English translation of the original, "Où sont les neiges d'antan?," rather than, say, the flat-footed "Where are the snows of previous years?" Moreover, and importantly for our purposes, recent developments in artificial natural-language processing suggest that artificial systems may be able to *learn* underlying syntactic and semantic structures of language from the task of developing compact, hierarchical, predictive, or generative models of large bodies of linguistic data and can use these models to guide interpretation.[10] Unlike systems that are preprogrammed with grammatical information, these systems use fairly generic learning methods to "acquire" from exposure to language data latent structures *in* language that then can be used for tasks like interpretation, translation, and similarity judgments for an open-ended array of sentences. Might similar kinds of general-purpose learning capacities enable artificial systems to extract from the context of human interaction ethically relevant latent structures of situations, actions, agents, and outcomes? And might this in fact be much closer to the way actual humans become sensitive to ethically relevant features or make ordinary, "intuitive" ethical judgments? This brings us to a developmental perspective on human cognitive and ethical capacities.

1.3. A Developmental Perspective

Recent years in developmental psychology have seen the emergence of learning-based, "constructivist" or "theory forming" approaches to cognitive phenomena previously attributed to specialized "innate modules."[11] For example, in the case of language learning, it has long been recognized that infants receive relatively little explicit instruction in language in their early years, yet during these years normally developing children acquire a remarkable degree of fluency in understanding and producing an open-ended array of novel sentences in their native tongue. Positing an innate, generative language module seemed to offer the only explanation of how this could occur, given the limitations of "associative learning" and the disproportion between the finite amount of language and language training to which children are typically exposed and the open-ended competence they acquire. This is sometimes known as the "poverty of the stimulus" argument.[12] But *is* the stimulus really impoverished? And is "associative learning" really so limited?

Since the heyday of innatism, we have learned a considerable amount from cognitive science about the potential for experience-based learning of rich, hierarchical structures, and from developmental psychology about the highly active experiential life of infants, even in their earliest days and weeks.[13] Moreover, innatism always faced the problem that infants must somehow already be able to detect many structural and contentful features of language in order to *apply* a category- and rule-based "innate grammar." After all, infants begin life in a complex, continuous acoustic environment within which they must learn to distinguish overheard *language* from other elements in the continuous stream of sounds and noise, and to attend to overheard language closely enough to detect patterns that permit the extraction of discrete, recurring units and combinations in the language, despite, for example, the wide variation in the acoustic profile of individual voices. Moreover infants need to be able to detect signs of adult attention and to track the intended referents of adult gestures or words. These are already formidable learning tasks in modeling structural features of the world the infant inhabits, and they must be solved for her specific acoustic and social environment. An innate grammar module on its own would not equip the infant to accomplish them. What could?

Recent developments in machine learning applied to natural language have begun to suggest how such learning might be possible through probabilistic means, even in the absence of much by way of explicit linguistic instruction.[14] Infants are, after all, exposed to a very large amount of overheard language and have at their disposal a very large amount of fast, flexible computational capacity. It seems they put both of these to good use. For example, we now have evidence that infants in the first weeks of life have begun to discriminate overheard speech

from the rest of their acoustic environment and are beginning to form calibrated expectations about phonetic regularities, which are manifest in greater surprise at, and interest in, novel or anomalous sequences of phonemes—a characteristic feature of probabilistic learning.[15] Over the course of the first year and a half, while an infant's explicit language capacity and adult explicit linguistic instruction are both typically limited, young infants have begun to piece together the social and intentional structure around them. By nine months they can discern others' goals on the basis of their behavior,[16] and by twelve to sixteen months they can relate means to goals,[17] follow others' attentional cues,[18] engage in joint attention,[19] and identify the intended referents of their words or gestures.[20] We see, then, a pattern of emerging competencies of a kind important for the development of language, accomplished gradually through the course of experience in a way that resembles their gradual learning of other kinds of causal relations and regularities in their world.

The stimulus infants receive, then, is not so impoverished after all, stretching over many months of observation of the behavior of persons and objects in their near vicinity. And probabilistic forms of learning turn this seemingly "passive" experience into more than "mere association." Instead it is a form of active *experimentation*, with the continuous formation of expectations on the basis of observed associations and continuous feedback from discrepancies between such expectations and actual outcomes. Since the physical and social world contain very significant structure, more effective and efficient prediction pushes infant learning in the direction of representing such structure, favoring the development of internal models that use abstraction and hierarchy to generalize projectively, without the need to posit an innate "language module."[21]

We can connect this idea of learning via experiential modeling to the child's challenge in moving from observation to action by reflecting on the so-called "Good Regulator Theorem" of control theory,[22] which holds that ideally effective and efficient regulation of a system requires the building and use in decision-making of a model of that system—a model representing the underlying structures and potentials of the system. Such a model can be used in a forward direction for intelligent simulation and action selection, and in an inverse direction for learning from subsequent experience. Models of this kind can also play a fundamental role in the development of motor control skills.[23]

Suppose, then, that we think of the infant mind as *regulating* its interactions with the environment, exercising whatever capacities it can to get its needs met. And no part of the infant's causal environment is more important for her than the *agents* in her life, so that causal and social learning are intimately linked, and intuitive psychology emerges alongside intuitive physics. It would, after all, be very difficult for the infant to build a predictive model of the world around her without taking into account the distinctive ways in which agents behave, and

beginning to model the "internal" as well as external sources of such behavior, much as infants begin to model latent as well as manifest causal relations.[24] Evidence suggests that infants develop piecewise an increasingly complex "theory of mind" or model of agents as continuing entities whose behavior is the product of perception, motivation, emotion, belief, and intention.[25]

Moreover, while the infant might start by using her own mind as a matrix for understanding others and their actions,[26] the pressure to develop more reliable expectations of others pushes in the direction of representing others' mental states in their own right—not as projections of the infant's own states. We see emerging in infants an ability to grasp that others may differ, first, in motivation, then in belief, then in perceptual knowledge, then in possessing false beliefs, and then in hidden emotions.[27]

Spatial representation in foraging animals (ourselves included) appears to involve the construction through experience of non-egocentric as well as egocentric maps.[28] These spatial representations can then be used to associate expected rewards with nonproximate locations and to simulate and compare possible pathways toward these rewards, facilitating more efficient and effective foraging.[29]

Likewise infant mapping of social space and its possibilities involves an ability to represent how things are in non-egocentric as well as egocentric terms, making possible more accurate, less position-dependent simulations of potential social interactions and evaluation of their likely outcomes. Over the course of the first years of life, when infants have only limited causal powers of their own, observation of others' actions and outcomes plays a fundamental role in the development of their own expectations and understanding.[30] The non-egocentric *epistemic evaluation* of others—observing others' interactions to map the reliability and competence of agents in their interactions with third parties—comes to play a critical role in shaping who infants are disposed to imitate or learn from, independent of personal affiliation.[31] As we will see, this ability to form and be guided by non-egocentric as well as egocentric representations and evaluations, which might be driven in the first instance by the need for accurate prediction, is of special interest for *ethical* development in children, since among the fundamental features of ethical evaluation are that it calls for an ability to represent non-egocentrically the nature and magnitude of the concerns of others, the likely results of one's own actions, the causal-intentional structure of others' actions, and whether others are reliable or trustworthy.

1.4. Default Trust and Default Cooperation

What are some of the characteristics of a developing psyche that would promote this kind of integrated learning about the causal world of things and agents?

Clearly, infants need to be motivated to attend carefully to experience and to notice patterns. They need to form expectations based upon such patterns, and to find failed expectations discomfiting in themselves, even when this does not directly touch their interests. And they need to respond to such anomalies by increasing their attention and effort, not by simply shrinking the scope of their expectations. This collection of features we can think of as *curiosity*, a form of internal motivation to learn above and beyond any more specific purpose the infant might have.

But curiosity is not enough. The brief description just given presupposes that infants are also disposed to *rely upon* or *trust* their own faculties—perception, association, memory, and so on—even without any guarantee of the reliability of these faculties. Without such a disposition toward *default reliance* or *default trust*, even an infant natively equipped with good eyes and ears and a keen mind would remain trapped in ignorance. After all, any evidence she might gather of the reliability or unreliability of her faculties would already depend upon the use of those faculties, in effect giving them some measure of default epistemic authority. Once some measure of default reliance or trust is in place, then the formation of expectations can begin to generate feedback from subsequent experience--a kind of bootstrapping. Bootstrapping does not mean *indefeasibility* however; indeed, default reliance and trust operate in the service of generating more determinate guesses, creating the potential for more informative errors and growing more selective or calibrated over time.

Somewhat metaphorically, we can think of such default, defeasible trust as a "prior" that enables the infant to *cooperate* with her faculties, by "playing" a cooperative move on the first turn by forming expectations as she would if her faculties were reliable, yet with no security that her faculties will prove cooperative in return by yielding reliable information. In contrast, for her to refuse to extend any unsecured cooperation to her faculties (in this metaphorical sense) would be a self-defeating epistemic strategy—not by incurring a risk of believing something false but by undermining the possibility of believing anything at all.

Consider now that portion of the infant's epistemic engagement with the world that is social, and where cooperation can be less metaphorical. Here too an infant initially disposed not to rely upon or trust those around her until she has confirmation that such reliance and trust will be well-placed would cut herself off from the very experiences she would need in order to learn whom or what to trust, and how much. As before, initial trust can be modulated by subsequent experience, so that expectations can become better calibrated to actual outcomes, and reliance and trust more selective.[32]

Infant default reliance and trust extend beyond the epistemic and can play a vital role in initiating cooperative relations with others. Early on, infant responsiveness *reinforces* adult attention, facilitating development of reliable channels

of communication between infants and caregivers that do not depend upon language. Infants are typically disposed to reciprocate care, to the extent that they can. For example, by the second year infants still crawling or toddling are able to form representations of adult goals from failed as well as successful adult behavior[33] and are spontaneously disposed to initiate an attempt to help an adult complete a failed task, even a stranger, and without encouragement or promised reward.[34] And toddlers who have participated in a successful shared task with a novel partner are spontaneously motivated to share the gains achieved by the task, again without explicit encouragement or reward.[35] These are manifest forms of a general disposition to default cooperativeness that has in fact been operative in the infant since early weeks of life, helping her to establish positive, reciprocal relations with those around her.

Infancy is an extreme case in which an individual's problem-solving capacities depend upon developing sustained, selective engagement with and reliance upon others. But as humans go through life, what they learn and what they are capable of achieving do not cease to depend extensively on coordination or cooperation with others. If anything, the scope of the coordination and cooperation with others needed for continued learning and success in attaining one's ends *grows* with time. For this to be sustainable, individuals must be motivated both to trust help from unrelated others and to help unrelated others in ways that reward their trust. As Hobbes pointed out over three hundred years ago, mutually beneficial cooperation among strangers is possible when individuals are disposed to initiate cooperation without requiring initial security (e.g., as a credible way of signaling willingness to cooperate) and to reciprocate cooperation when it is received.[36] More recently, game theorists have shown that this set of dispositions can become widespread within a population and be effective in resisting "invasion" by more opportunistic agents.[37] And a large-scale survey of hunter-gatherer societies suggests that a capacity for coordination and cooperation with others, including nonkin, mediated by forms of reciprocity that are indirect and temporally extended, may play a central role in explaining how human hunter-gatherers have succeeded over millennia in maintaining egalitarian social cohesion in the face of limited resources, without the forms of dominance hierarchy found in the great apes.[38] Recent research suggests that the disposition to give weight to the interests of others that is not simply mediated by one's own interests is something like the default stance of ordinary human interaction and can be self-reinforcing[39]—including, one might stress, human communication and information exchange, as the norms of conversation attest.

So much of what we think of as an individual human's general human intelligence or problem-solving capacity is really social in origin, character, or operation that we should think of the ability to initiate and sustain productive social connectedness with others as an additional basic faculty of learning, which

supplements other basic faculties like perception, memory, and reasoning. Default cooperative dispositions with respect to others therefore are as much a part of the human capacity for learning as default cooperative dispositions with respect to one's own basic mental faculties.

Language is fundamental to general intelligence and problem-solving capacities typical of humans, which are able to draw upon social knowledge built up over generations of experience. And language is an outstanding example of what default cooperative dispositions among nonkin can accomplish for any species that can achieve them. A shared language can be sustained only because enough speakers regularly use the language with sincere and helpful communicative intent to make it worthwhile for us to speak with each other and rely upon what each other says—to make openness to conversational exchange, overall, a positive-sum activity. Open conversational exchange among strangers is a form of mutual constraint and contribution for mutual benefit, and it plays an essential role in knitting together and facilitating the large-scale forms of cooperation and accommodation upon which human culture depends.

Individual human intelligence and problem-solving ability at age two is said to be quite comparable to that of a chimpanzee of the same age. But human two-year-olds are able to do something even adult chimpanzees are not, and that is fundamental to the extensive growth of human intelligence and problem-solving ability: to come together spontaneously with others to accomplish a task requiring joint attention and coordinated playing of understood roles, and, equally spontaneously, to share the rewards of cooperation with others without further incentive.[40] The divergence in cognitive accomplishment and practical problem-solving that comes as humans work together—the emergence of shared languages, of extensive forms of social learning, culture, and exchange—explains why *Homo sapiens* could overrun the planet, making their own habitats as needed, while *Pan troglodytes* is at risk of disappearing from the wild as its natural habitat shrinks.

1.5. Ethical Development and Ethical Judgment

We have spent so much time on the questions about the capacities underlying aspects of language and epistemic development because they afford us insight into the capacities that underlie ethical development as well. It is no accident that the norms of conversation, for example—of mutual recognition, of according others some authority to contribute, of seeking to determine the communicative intent of others and signaling this to them, of seeking to reply in ways that could be comprehensible, relevant, and responsive to others' concerns, and so on—are so close to norms for productive epistemic exchange. Now we will add: it

is no accident that they are also so close to norms for ethical interaction. Indeed a large and influential tradition in ethics, *communicative ethics*, is built around this fact.[41]

Intriguingly, the step-wise development of children's ability to model others' minds predicts a range of features of infant behavior that have strong relevance to ethical learning. For example, even controlling for other abilities, a child's development of theory of mind is predictive of her current and future *maturity*, as manifest in the ability to form positive relations with peers. Such abilities include: understanding the needs and interests of others, even when different from oneself or one's group; standing up for one's own opinions, needs, and rights; successfully joining new groups or welcoming new members into one's own group; playing or working together with peers without conflict; and coping with conflicts that do arise.[42] These are all skills that involve apt responsiveness to ethically relevant features, as understood by virtually any widely held ethical theory. Just as there was a parallelism in the development of causal understanding and theory of mind, there is a parallelism in the development of theory of mind and capacity to be aptly responsive to ethically relevant features.

For example, assessment of *intent* is a core component of understanding the causal, epistemic, and ethical character of an action, so acquiring the ability to distinguish intentional from unintentional actions is important for prediction (e.g., what to expect next), learning (e.g., whether an adult error was the result of ignorance or is a sign of unreliability), and ethical assessment (e.g., whether a harmful action by an individual was an accident or is a sign of ill will or untrustworthiness). By the end of their first year infants have begun to use situational cues to determine whether an action is intentional to modulate their responses in all three domains.[43]

Ethical development appears to begin earlier than the explicit inculcation of social norms by adults and also to develop in ways that are both more basic—for example, in grasping what behavior, in a given context, constitutes a harm—and more autonomous than external instruction. An example of autonomy, mentioned earlier, is the fact that three- and four-year-olds across a range of cultures show a spontaneous ability to question rules given to them by figures in authority, and will resist following a rule given by a figure in authority if they see this rule as unduly harmful or unfair. Moreover, they will cite these ethically relevant features to explain their resistance.[44] At the same age, children will spontaneously share their gains from a joint activity with a co-participant to redress an unfair distribution or unwarranted punishment by a figure in authority,[45] and will spontaneously attempt to stop third-party ethical transgressions.[46] Just as infants are to a considerable degree autonomous, experience-based causal learners[47] and learners of theory of mind,[48] capable of forming without explicit

instruction non-egocentric representations and evaluations of their causal, social, and epistemic environment, so infants appear to a considerable extent to be autonomous, experience-based ethical learners, capable of forming without explicit instruction the kinds of non-egocentric representations and evaluations of situations, agents, actions, and outcomes upon which responsiveness to ethically relevant features is based—as manifest, for example, in the social skills and maturity mentioned earlier. There is a dark side to such socially oriented learning: as infants gain in sophistication about social relations, they become more oriented toward what they find familiar or, somewhat later, toward people they perceive as members of their own group. However, while debate persists on this question, such "own group" preference does not appear to be a "wired-in" response as such and does not prevent infants or adults from being capable of an extraordinary degree of spontaneous cooperation and collaboration with unrelated individuals, especially in comparison with our nearest animal relatives.[49] And social learning in settings involving shared activities and goals can help counteract implicit bias.[50]

But what if we look beyond the developmental setting? What evidence do we have of the kinds of capacities that could underlie the ethical judgments of adults? Here we will briefly consider two kinds of evidence, from neuroimaging studies of ethical judgment and from informal classroom sampling of "ethical intuitions."

The question of the neural basis for ethical judgments has generated a large volume of research, the general trend of which has only fairly recently become clear. Initially, partly under the influence of innatist notions of a "moral module," it was thought that there might be some region or regions of the brain specialized for ethical judgment. By contrast, the approach to ethical development sketched here would predict that the neural substrate of ethical judgment would involve regions or networks subserving general-purpose learning and judgment concerning a range of causal and theory-of-mind–related questions about situations, actions, outcomes, and agents. Recently, metastudies of experimental reports of neural imaging during ethical judgment have come to the conclusion that ethical judgment relies heavily upon just such a neural network of regions, the *default network*.

The "default mode" of brain functioning is one of two primary modes of brain activity, alternating with the more focused "attentional mode."[51] Each mode corresponds to higher levels of coordinated activation in a relatively stable, interconnected set of brain regions. What are someof the functions of default network processing? They include, most importantly, episodic and semantic memory, scene construction and the imaginative simulation of possible futures, counterfactual reasoning, inferring the mental states of others, self-referential processing, and ethical judgment.[52] In other words, the primary network subserving

ethical judgment has the features that would be predicted by a model of ethical development as continuous with these other forms of cognition and evaluation.

There are of course many complexities and pitfalls in any appeal to neuroimaging evidence, and we can distinguish multiple kinds of ethical judgment, such as active versus passive, self-referring versus other-referring, and intuitive versus deliberative.[53] It is therefore still much too early to have any definitive picture of the neural basis of ethical thought and feeling. But neuroimaging using a variety of techniques has thus far been largely consistent with the idea that ethical cognition is supported by domain-general processing and essentially continuous with other ways in which we size up situations and actions and make evaluations and choices.[54]

More broadly, neuroimaging and connectivity research have increasing put in question the kind of "affective versus cognitive" division of mental processing found in many "dual-process" models of ethical cognition.[55] There are indeed forms of processing located in regions of the brain associated with affect that interact early and quickly with sensory input, before higher-order declarative reasoning has begun to operate, but these are also systems that subserve probabilistic learning, spatial mapping, evaluative comparison, and other core elements of "cognition."[56] Increasingly, a picture is emerging of cognition as widely distributed in the brain, and the age-old idea of the mind as pitting "reason" against "emotion" may be an artifact of our limited insight into the ways in which our minds actually operate. "Affect," as psychologists understand it, is not simply a matter of aroused emotion but is a capacity of the brain to synthesize multiple streams of information and evaluation in a manner that can orient or reorient a suite of mental processes—attention, perception, memory, inference, motivation, action-readiness—in a coordinated way to address actual or anticipated challenges.[57] If we are asking how an artificial system might make intelligent decisions responsive to ethically relevant features, we may wish to emulate the functional characteristics of this design,[58] which is inherited from our animal ancestors and highly conserved evolutionarily.

"Ethical intuitions" have also been subject to extensive research in recent decades. "Intuition" here does not designate a specific kind of mental process as such, but rather an assessment—whether of a particular scenario, a type of action, or a general principle—that is often relatively fast and effortless yet that typically feels compelling even though we have little insight into the process by which we arrived at it and may be unable to articulate a satisfactory rationale for it.

A principal focus of discussions of ethical intuition in recent decades, and of discussions of ethics and artificial intelligence as well, has been the "Trolley Problem," a puzzling pattern of ethical intuitions reliably evoked by a series of scenarios involving runaway trolleys. Trolley problems have sometimes been

called the *Drosophila* of ethical inquiry—a shared, heavily studied "test bed" for hypotheses about ethical judgment. It is, moreover, a nice irony that trolley problems, long castigated by critics as hopelessly artificial, turn out to have such direct analogues in one of the most important actual applications of artificial intelligence to life-affecting decision-making to date: self-driving vehicles. Let us begin, then, to look at trolley problems to see whether we can discover anything of relevance to our discussion of the nature and origin of human responsiveness to ethically relevant features. I believe we can.

To make my argument, I will be drawing upon in-class, confidential sampling of the intuitive ethical judgments of undergraduates in large ethics lectures I have taught at the University of Michigan over a number of years. During lecture, students are able to respond rapidly and confidentially to questions I pose by using individual wireless keypads (iClickers) that transmit their responses to a receiver at the front of the room. I am then able to display the overall patterns of response on a screen for students to see. These are hardly controlled experiments, and so they must be considered suggestive only.[59] But their informality also has advantages, in that it enables me to push a bit beyond the usual tightly constrained diet of standard examples in the trolley literature, and perhaps to probe a bit beneath the surface of my students' responses.

I needn't here rehearse the particulars of the most familiar trolley problems, which we will call "Switch" and "Footbridge."[60] In their responses to Switch and Footbridge, my students typically exhibit the same pattern of response that has been found repeatedly in the literature. In Switch, a strong majority (typically about 80%) say that one *should* push a lever to switch the runaway trolley to a sidetrack, saving five workers down the main track but killing one worker on the sidetrack. And in Footbridge, a strong majority (typically about 75%) says that one *should not* push a large man off a footbridge to stop the runaway trolley to save five workers. Despite a certain abstract similarity of the two scenarios—in both, an intervention taken to prevent the deaths of the five workers on the main track brings about the death of one other individual who is not initially at risk—the asymmetry in intuitive judgment has proven remarkably robust. Even moral philosophers who have considered the problem for years and who themselves judge that one *should* push the man off the footbridge tend to admit that this scenario does not cease to trouble them. Since the trolley problems first emerged in the 1970s, dramatic changes have occurred in people's views about interracial marriage, women's roles, gay marriage, premarital sex, smoking marijuana, and more. Yet the trolley problem asymmetry remains pretty much undiminished. Further, the asymmetry has been found cross-culturally, doesn't manifest gender differences, and appears both in vivid virtual-reality simulations and in simple, undramatic, verbal posing of the dilemmas.[61]

The problem continues to fascinate because there has been no analysis of the asymmetry that has received wide acceptance, despite many attempts.[62] One promising early explanation—roughly, that in Footbridge one is deliberately using the worker killed as a "mere means," whereas the worker dies in Switch as an "unintended side effect"—has lost adherents owing to a case called "Loop": Suppose the switch could send the trolley down a sidetrack that loops back to the main track; however, this will stop the trolley from hitting the five workers because a single, large worker is currently on the sidetrack, and the trolley, hitting him, will stop before rejoining the main track. Should you switch the trolley, killing one in order to save five? Here, according to the standard interpretation, the single worker on the sidetrack is being used as a "means" in essentially the same way as "Footbridge," since his being struck by the trolley is not an unintended side effect but essential to saving the five. Despite this, a strong majority of my students (typically about 80%), and of most populations sampled, say one *should* push the lever to send the trolley onto the looping sidetrack.[63]

At this point "dual-process" psychologists entered the fray, arguing that the difference between Switch and Loop, on the one hand, and Footbridge, on the other, is attributable not to a matter of ethical principle but to a rapid, strong, automatic, affectively charged, negative System 1 (or, more recently, "model-free") reaction to the thought of using direct muscular force to kill the man in Footbridge. This rapid "push button" response does not occur in cases like Switch and Loop, where the victim is less proximate and one's effect upon the victim less direct, so that the System 1 response is relatively weak, and a slower and more deliberative System 2 (or "model-based") response can come into play, favoring a calculation of minimizing harm.[64] This dual-process account affords an explanation of the asymmetry, but not one that provides much by way of ethical justification. Hence, some have argued on this basis that we should discount the normative significance of the Footbridge verdicts.[65]

But now consider "Beckon": As before, the runaway trolley will strike and kill five workers if not stopped. You are at some distance from the track, with no access to a switch, but you see a large man standing on the other side of the track, facing in your direction but unable to see the trolley approaching. If you conspicuously beckon to the man, encouraging him vigorously to come in your direction, he will step onto the track and immediately be struck and killed by the trolley, stopping it before it hits the five workers. In classroom sampling, I have regularly found that 60% to 70% of students say that one *should not* beckon to the man. This despite the fact that the death he suffers happens at a distance and involves no direct exertion of my own muscular force upon him.

Is intentionally gesturing in a way that lures someone to his death the problem? Consider now "Wave": You are standing down the track from the five workers, who are looking in your direction and do not see the trolley approaching them

from behind. If you wave vigorously to the side, encouraging them to step in that direction, the five workers will step off the track and be saved. However, another worker who is looking your way and who is initially standing *alongside* the track will also see your waving gesture and step in the same direction. This will place him on the track, where he will be struck from behind and killed. Here some 70% to 90% of my students will say that one *should* wave to the five workers, saving them but killing the one man lured thereby onto the track. What, then, could explain *this* asymmetry, which is as pronounced as the original Switch versus Footbridge asymmetry?

Suppose for argument that the earlier account of ethical judgment—as involving general capacities for modeling, simulating, and evaluating situations, actions, agents, and outcomes—was accepted. This would suggest that we should look for a complex competence in understanding the social landscape and its possibilities underlying all these trolley problems. How might we find out? When I ask my students whether learning that their roommate has been in a Switch-like trolley problem and has pulled the lever to send the trolley down the sidetrack would increase, decrease, or not affect the *trust* they have in their roommate, the majority response typically is "no change in trust," while "increase trust" and "decrease trust" each receives a smaller number of votes. When a similar question about trust is asked about a roommate who took action in Loop, student answers are essentially the same. But when asked about learning that a roommate has pushed a large man off a footbridge in a Footbridge scenario, the strong majority response (typically 70% to 80%) is "decrease trust," with a much smaller number indicating "no change" and virtually no one indicating "increase trust." In fact, in a typical sample, a much smaller number indicate "increase trust" in the Footbridge case (around 5%) than had originally judged that one *should* push the man (around 25-30%).

So now, what about Wave and Beckon? Here the response in Wave is essentially indistinguishable from that in Switch and Loop, while the response in Beckon is essentially indistinguishable from that in Footbridge. As in Footbridge, a smaller number indicate "increase trust" (about 5%) than initially judged one should take the action in question (about 35%). This pattern of trust judgments has been found each year I've sampled my students, reliably grouping Switch, Loop, and Wave into one category with regard to trust, and Footbridge and Beckon into another.

Perhaps, then, my students' intuitive responses to individual trolley scenarios involves not simply thinking about the *act* involved but thinking "What kind of person would perform this act?" and perhaps "Would I?" Personality tests have been given to subjects about to be given trolley problems, and those giving a "push" response in Footbridge as a group, in comparison to the group giving a "don't push" response, scored on average higher on psychopathy scales

and higher in indifference to harm or to ethical violations generally, while they scored lower on perspective-taking and altruism.[66] It would seem that my students' trustworthiness judgments may be tracking something real about "the kind of person who would perform this act."

But why would this consideration show up in an intuitive sense of what one should do in a given scenario? Suppose, as *virtue theorists* such as Aristotle[67] and Hume[68] have argued, our primary access to our ethical understanding is not via highly general principles or judgments of particular acts, but via our general sense of the tendencies of certain kinds of traits of character or motivational structures. Looked at from a modeling perspective, one might think one gains greater predictive and explanatory purchase in ethical thought if one assesses those around one in terms of their general dispositions to act or their trustworthiness. To gain an idea of how to act in a given situation, then, it may be more reliable to ask whether someone who manifested skills and traits of character we'd ethically admire would perform the act.

To further examine this interpretation, I ask my students what emotions they would expect to feel, had they intervened in a trolley problem to save the five and afterward decided to approach the family of the single victim they had killed. In the case of Switch, Loop, and Wave, the predominant response is to anticipate feeling regret and *guilt*, with some expectation that the family might understand. In the cases of Footbridge and Beckon, the predominant response is to anticipate feeling regret and *shame*, with little or no expectation that the family would understand. Anticipated shame, as opposed to anticipated guilt, suggests that they think others would also think that performing the interventions in Footbridge and Beckon would be a sign of defective character.

My hypothesis is that, when making an ethical assessment, my students (and the rest of us) rely upon acquired, general, abstract causal-evaluative models of situations and agents to simulate possible actions and likely outcomes or reactions. The simulations can be quite complex: *How would it feel to perform this action? Could I actually see myself doing it? What kind of person would perform it? What would others think, and could I face them?* But this kind of real-time simulation and evaluation of possibilities, and associated feelings and reactions on the part of others is exactly the kind of *prospective* processing the human default system appears to be engaged in systematically, off and on throughout the day, as we navigate the physical and social environment.[69]

This picture of intuitive ethical judgment also fits the recent proposal that prospection is a fundamental organizing principle of the human brain.[70] And it echoes the idea that prediction is of the essence in learning and intelligence, whether animal or machine. Relatedly, several recent studies of ethical judgments[71] have found that a model of the hypothetical agent and choice seems to mediate the "intuitive" judgment of the action.[72] If some elements of people's

acquired causal-evaluative models of situations and agents are based upon extensive, ordinary experience of a kind most people could be expected to have, whatever their social or cultural identity, this would explain how some patterns of intuitive ethical judgment, such as the trolley asymmetries, could be found very widely and remain stable across a number of social or cultural changes. It would also help explain why the source of such patterns might be difficult to introspect, and why the patterns might nonetheless remain confident even though they cannot be fit to a priori ethical principles.[73]

This brings us to the "realistic Trolley Problem" that has been much discussed in connection with self-driving cars. I have polled my students about two possible <situation, action> rules that might be "programmed into" self-driving cars with one passenger aboard: (1) they might be programmed to swerve to avoid five individuals in a cross-walk, even in cases where this would result in the death of one other individual, not now at risk, on a side walkway; (2) they might be programmed to swerve to avoid five individuals in a cross-walk, even in cases where this would result in colliding with a concrete wall, killing the occupant in the car. When I first posed these questions to students several years ago, a strong majority agreed with programming (1) but disagreed with programming (2). This initially seemed to replicate the kind of asymmetry found in Switch versus Footbridge or Wave versus Beckon. However, over the intervening years the percentage approving programming (1) has remained consistently high, at 70% to 80%, but the percentage approving programming (2) has climbed from 35% to 65%.

Why has the original asymmetry not been robust? One potential explanation: these are cases in which the *agent* and questions about the *character* of the agent have been removed from the situation. Initially students might have been tempted to assimilate self-driving cars to personified agents, but as the problem of regulating self-driving cars became more familiar over time, students became more likely to think of the problem in terms of *general rules*, and from that standpoint, there seems to be no reason to assign special weight to the car's occupant over pedestrians. Interestingly, the initial asymmetry actually went away over the course of the term, as discussion of the case proceeded. By contrast, the initial trolley asymmetry, and that between Wave and Beckon, tend to persist from one end of the term to the other, despite extensive discussion.

1.6. Artificial Ethical Psychology

The burden of the argument thus far is that we should understand the human capacity to identify and respond to ethically relevant considerations as an integral part of the competencies and knowledge we acquire that underwrite

human general intelligence and capacity for open-ended problem-solving. The reasoning has drawn heavily on evidence from human psychology, but many elements of the argument do not turn on details specific to *Homo sapiens*. For example, the ways in which default, defeasible trust or cooperation can make possible positive-sum results in learning, language, and social interactions depend upon very generic dynamics of agents and groups of agents.

Perhaps our model of how to develop *machine ethics* should not be based upon the idea of "programming in" principles or designing machines to "align themselves" with the preferences or values of the humans they encounter. Neither of these seems to be the way in which humans acquire ethical competence. It is not primarily "inculcated" into children by explicit adult teaching; indeed in many societies there is relatively little direct instruction of children. And, as we have seen, children display greater autonomy than simply aligning themselves with the preferences or values of the adults around them. Humans are hardly ideals of ethical competence, but if we wish to develop machines at least as trustworthy with life-affecting decision-making as ordinary humans, perhaps we should look to models of the development of machine ethics that more closely approximate human ethical learning.

In truth, we do not know what the principles would be for "programming in" ethics as anything like an operational system. There is continuing disagreement over the fundamental principles of ethics, and even supposing this were not so, there is sufficient distance between fundamental principles and actual applications (*What constitutes a harm in a given instance? How to assess the relative magnitude of harms and benefits? When have the conditions of a promise been sufficiently undermined that it no longer binds?*) that a large quantity of ethical understanding is needed in order to apply them—understanding that seems to come only with extensive individual and shared experience and is not contained within the principles themselves. Even if we consider a disaggregated system of less fundamental ethical rules, the actual situations we encounter are too varied, and the kinds of considerations that need to be taken into account too diverse, to allow these rules to be more than rules of thumb. Perhaps the ethical theory closest to common sense in contemporary Western society is W. D. Ross's system of prima facie duties,[74] but it is a fundamental feature of Ross's account that these duties can come into conflict and that there are no strict rules for determining which duties are weightier in a given case. Instead, Ross argued, we must have recourse to *intuition*.

It is a striking fact about ethical judgment that it has such a strong intuitive element, even in the assessment of ethical theories. What might a model of our ethical competence look like that would make sense of this idea of intuition? While some philosophers have thought of intuition as something like direct rational perception of self-evident truths, a long tradition in philosophy holds that

intuition is more like common sense—a large body of understanding, relying heavily upon experience and social discussion, without the structure of a deductive system, yet carrying enough structured information to make nuanced judgment possible. Fortunately, recent research in artificial intelligence is beginning to give us an idea of what such a large body of intuitive understanding or common sense might look like, how it might be acquired through experience, and how it might support abstract generalizations as well as nuanced judgments of novel cases, despite lacking an overall deductive, rule-like structure.[75] Rather than rely upon preprogrammed feature detectors or policies, programs that have been successful in tasks such as image identification, natural-language processing and translation, game playing, and motor control have been able to acquire high levels of competence via processes of learning based upon autonomous development of complex, generative representations of large bodies of data.

Here we have speculated that, in the case of humans, a relatively modest set of priors—for example, curiosity and default, defeasible reliance and trust—could combine with basic faculties and ample learning capacity to promote the acquisition of predictive and generative representations of the physical and social world, for example, intuitive physics, intuitive psychology, or communicative competence. We speculated further that these same capacities can subserve epistemic and ethical evaluation, and gave some evidence from neuroscience and ordinary ethical judgment to support this speculation. Thinking about machine ethics may need to undergo the same kind of "learning revolution" that thinking about machine learning and expertise, and thinking in developmental psychology, have undergone in recent years.

How might artificial ethical learning proceed? Here I have no expertise. Fortunately, however, this question is already being examined under other descriptions. For example, machines that are learning to carry on effective natural-language conversations—for example, to provide customer service that satisfactorily identifies and addresses problems, or to provide companionship and reliable health monitoring for an elderly person living at home, or to help students learn by identifying their strengths and weaknesses and drawing upon their abilities and motivations—are acquiring skills in understanding people and their needs and aims, and learning what it is like to work together with people to achieve mutually desirable outcomes. As artificial systems are increasingly deployed in our lives, the *social* dimension of their existence—their ability to work together with humans and together with one another—will become increasingly important and a fundamental part of the enlargement of their intelligence. We should be asking how systems, equipped with such priors as curiosity and default, defeasible trust or cooperativeness, might come to be themselves complex social agents, subject to demands that require them to figure out how to achieve solutions by working with others, sharing out tasks and assigning

responsibility in ways that can achieve positive sums and help sustain further cooperation—*learning* fundamental elements of ethics and knitting them into their global knowledge and competence.

Just as infants observe countless hours of adult behavior seeking to predict what will happen next, machines can observe countless hours of human and machine behavior seeking to predict, first, the next instant, then, the next second, then the next minute, and so on. They can learn to read the goals and beliefs of those around them, learning, as "mature" children do, such skills as recognizing the needs of those around them, even those who depart from the norm, or standing up for their own interests while according weight to the interests of others, or entering into novel relations and helping others to do so without conflict, and so on.[76] Adversarial training might pit machines observing human interactions and making predictions, or telling credible stories, against humans doing the same, asking the discriminator to determine whether the source is artificial or human. Machines can be apprentices or partners in complex tasks, learning social as well as technical skills. Self-driving cars, for example, can learn how to manage the elaborate interactions involved in a crowded parking lot at holiday time, or merging into bridge traffic at rush hour, in ways that achieve a successful mix of forcefulness and deference, reading the intentions of other drivers or autonomous vehicles in order to find workable solutions. Like humans, machines can use their internal models to create non-egocentric as well as egocentric representations and evaluations of situations, actions, outcomes, and policies. Like humans, machines can use these models to maintain some degree of autonomy in evaluation and action. We already know that machines should not be built so that they will pursue whatever goal they are given unquestioningly. Intelligent machines, like intelligent animals, should operate with modulated uncertainty rather than absolute certainty, and should be able to use their own resources, and draw upon others as a resource, for criticism and self-criticism.

Human learning is most impressive when it leverages the ability to form communicative and cooperative relations with others that extend our problem-solving capacity far beyond whatever we individuals could accomplish on our own. Artificial learning likewise can reach its fullest development socially and cooperatively, drawing upon an expanding network of perspectives and experience. The threat of an emergent "superintelligence" or, much more proximately, of artificial intelligence working in the service of those who'd rather dominate and exploit than work together and share, can only be met by developing a sufficiently robust community of cooperating human and artificial intelligences that takes advantage of the fact that a society capable of joint effort and sharing is in the long run likely to know more, and adapt more readily and with greater foresight, than a society based upon subordinating the interests of the many to the

interests of the few, and the suppression of alternative points of view. We can only speculate, but from the perspective of learning, it would seem that humans are more valuable as cooperation partners than as peons or fodder. Indeed superintelligent machines themselves should be able to see this, especially if we've had the sense to enable them to grow up socially, as our partners in learning.[77]

Notes

1. The use of the terms 'ethical' and 'moral' often gives rise to puzzlement over the relation between the two. In empirical psychology, where emphasis tends to be placed upon the person and personality, the terms 'moral' and 'morality' are most often used—e.g., "moral development" and "moral judgment". By contrast, in many areas of normative application the terms 'ethical'and 'ethics' are most often used, e.g., "medical ethics" and "ethics and artificial intelligence". In philosophy, and in this paper, these terms are used almost interchangeably, though 'ethical' often includes *prudence* as well as morality proper. E. Turiel, *The Culture of Morality: Social Development, Context, and Conflict* (Cambridge: Cambridge University Press, 2002); J. G. Smetana et al., "Developmental Changes and Differences in Young Children's Moral Judgments," *Child Development* 83 (2012): 683–96.
2. R. Frank, *Passions within Reason: The Strategic Role of the Emotions.* (New York: W.W. Norton, 1989); J. Decety and A. N. Meltzoff, "Empathy, Imitation, and the Social Brain," in *Empathy: Philosophical and Psychological Perspectives*, ed. A. Copland and P. Goldie (New York: Oxford University Press, 2011), 58–81.
3. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
4. Oderbank, S.G., J. Nitschke, A. Mokros, E. Habermeyer, and O. Wilhelm, "Psychopathic Men: Deficits in General Mental Ability, Not Emotion Perception," *Journal of Abnormal Psychology* 127 (2018): 294-304; Kavish, N., C. Bailey, C. Sharp, and A. Venta, "On the Relation between General Intelligence and Psychopathic Traits: An Examination of Inpatient Adolescents," *Child Psychiatry and Human Development* 49 (2018): 341-51.
5. R. J. R. Blair, "The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy," *Trends in Cognitive Sciences* 11 (2007): 387–92; R. J. R. Blair, "The Emergence of Psychopathy: Implications for the Neurophysiological Approach to Developmental Disorders," *Cognition* 101 (2006): 414–42.
6. Monagan, C., H. Bizumic, and M. Sellbom, "Nomological Network of Two-Dimensional Machiavellianism," *Personality and Individual Differences* 130 (2018): 161-72.
7. Bereczkei, T., P. Papp, P. Kinces, B. Bodrogi, G. Perlaki, G. Orsi, and A. Deak, "The Neural Basis of the Machiavellians' Decision Making in Fair and Unfair Situations," *Brain and Cognition* 98 (2015): 53-64.

8. Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. *The Foundations of Human Sociality: Economic Experiments and Ethnography Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press, 2004.
9. M. Hoffman, *Empathy and Moral Development: Implications for Caring and Justice* (Cambridge: Cambridge University Press, 2001); Decety and Meltzoff, "Empathy, Imitation, and the Social Brain."
10. A. Van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv*, last revised January 22, 2019, arXiv:1807.03748v1; J. Devlin, M.-W. Chang, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, last revised May 24, 2019, arXiv:1810.04805v1,https://arxiv.org/abs/1810.04805.
11. S. Pinker, *The Language Instinct* (New York: Morrow, 1994); E. S. Spelke and K. D. Kinzler, "Innateness, Learning, and Rationality," *Child Development Perspectives* 3 (2009): 96–98.
12. G. K. Pullum and B. C. Scholz, "Empirical Assessment of Stimulus Poverty Arguments," *Linguistic Review* 19 (2002): 9–50.
13. J. B. Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science* 331 (2011): 1279–85.
14. Van den Oord, Li, and Vinyals, "Representation Learning with Contrastive Predictive Coding"; Devlin, Chang, and Toutanova, "BERT."
15. R. N. Aslin, J. R. Saffran, and E. L. Newport, "Computation of Conditional Probability Statistics by 8-Month-Old Infants," *Psychological Science* 9 (1998): 321–24; C. Kidd, S. T. Piantadosi, and R. N. Aslin, "The Goldilocks Effect: Human Infants Allocate Attention to Sequences That Are Neither Too Simple nor Too Complex," *PLOS-One* 7 (2012): e36399.
16. T. Behne et al., "Unwilling versus Unable: Infants' Understanding of Intentional Action," *Development Psychology* 41 (2005): 328–37.
17. M. Carpenter, J. Call, and M. Tomasello, "Twelve- and 18-Month-Olds Copy Actions in Terms of Goals," *Developmental Science* 8 (2005): F13–F20.
18. J. Moll and M. Tomasello, "12- and 18-Month-Olds Follow Gaze to Spaces behind Barriers," *Developmental Science* 7 (2004): F1–F9.
19. F. Warneken, and M. Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees," *Science* 311 (2006): 1301–3.
20. J. Halberda, "The Development of a Word-Learning Strategy," *Cognition* 87 (2003): B23–B34.
21. Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (2011): 1279–85; H. M. Wellman, *Making Minds: How Theory of Mind Develops* (Oxford: Oxford University Press, 2014); Goodman, N. D., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths. "A Rational Analysis of Rule-Based Concept Learning." *Cognitive Science* 32 (2008): 108–54. Goodman, N.D., M.C. Frank, T.L. Griffiths, J.B. Tenenbaum, P.W. Battaglia, and J.B. Hamrick, "Relevant and Robust: A Response to Marcus and Davis," *Psychological Science* 26 (2015): 539-41.
22. R. C. Conant, and W. R. Ashby, "Every Good Regulator of a System Must Be a Model of That System," *International Journal of Systems Science* 1 (1970): 89–97.

23. E. Todorov and Z. Ghahramani, "Unsupervised Learning of Sensory-Motor Primitives," *Proceedings of the 25th Annual International Conference of the IEEE EMBS* (2003): 1750–53; E. Todorov, "Optimality Principles in Sensorimotor Control," *Nature Neuroscience* 7 (2004): 907–15; O.-S. Kwon and D. C. Knill, "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning," *PNAS* 110, no. 11 (2013): E1064–E1073, https://doi/10.1073/pnas.1214869110.
24. A. Gopnik and H. Wellman, "Reconstructing Constructivism: Causal Models, Bayesian Learning, and the Theory Theory," *Psychological Bulletin* 128 (2012): 1085–108.
25. Wellman, *Making Minds*.
26. A. N. Meltzoff, "'Like Me': A Foundation for Social Cognition," *Developmental Science* 10 (2007): 126–34; J. N. Saby, A. N. Meltzoff, and P. J. Marshall, "Infant's Somatotopic Neural Responses to Seeing Human Actions: I've Got You under My Skin," *PLOS ONE* 8, no. 10 (2013): e77905, https://doi:10.1371/journal.pone.0077905.
27. H. M. Wellman and D. Liu, "Scaling of Theory-of-Mind Tasks," *Child Development* 75 (2004): 523–41.
28. E. I. Moser, E. Kropff, and M.-B. Moser, "Place Cells, Grid Cells, and the Brain's Spatial Representation System," *Annual Review of Neuroscience* 31 (2008): 69–89.
29. A. Johnson, M. A. A. van der Meer, and A. D. Redish, "Integrating Hippocampus and Striatum in Decision-Making," *Current Opinion in Neurobiology* 17 (2007): 692–97; A. S. Gupta et al., "Hippocampal Replay Is Not a Simple Function of Experience," *Neuron* 65 (2010): 695–705; A. D. Redish, "Vicarious Trial and Error," *Nature Reviews: Neuroscience* 17 (2016): 147–59.
30. A. N. Meltzoff et al., "Foundations for a New Science of Learning," *Science* 325 (2009): 284–88.
31. M. A. Koenig, V. Tiberius, and K. Hamlin, "Children's Judgments of Epistemic and Moral Agents: From Situations to Intentions," unpublished manuscript.
32. Ibid.; though on the role of native predispositions versus learned preferences in early filial responses, see E. Di Giorgio et al., "Filial Responses as Predisposed and Learned Preferences: Early Attachment in Chicks and Babies," *Behavioural and Brain Research* 325 (2017): 90–104.
33. H. Gweon and L. Schulz, "16-Month-Olds Rationally Infer Causes of Failed Actions," *Science* 332 (2011): 1524.
34. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees"; R. Roth-Hanania, M. Davidov, and C. Zhan-Waxler, "Empathy Development from 8 to 16 Months: Early Signs of Concern for Others," *Infant Behavior and Development* 34 (2011): 447–58.
35. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees."
36. Thomas Hobbes, *Leviathan* (1651), ed. C. B. MacPherson (London: Penguin, 1968).
37. R. Axelrod and D. Dion, "The Further Evolution of Cooperation," *Science* 242 (1988): 1385–90.
38. C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame* (New York: Basic Books,. 2012).

39. J. K. Rilling et al., "A Neural Basis for Social Cooperation," *Neuron* 36 (2002): 395–406; D. G. Rand, J. D. Greene, and M. A. Nowak, "Spontaneous Giving and Calculated Greed," *Nature* 489 (2012): 427–30; M. Crockett et al., "Harm to Others Outweighs Harm to Self in Moral Decision Making," *PNAS* 111 (2014): 17320–25; O. FedlmanHall et al., "Empathic Concern Drives Costly Altruism," *NeuroImage* 105 (2015): 347–56.
40. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees."
41. S. Benhabib and F. Dallmayr, eds., *The Communicative Ethics Controversy* (Cambridge, MA: MIT Press, 1990).
42. C. Peterson et al., "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development," *Developmental Psychology* 52 (2016): 46–57.
43. Koenig, Tiberius, and Hamlin, "Children's Judgments of Epistemic and Moral Agents," Unpublished manuscript (2019)
44. Turiel, *The Culture of Morality*; Smetana et al., "Developmental Changes and Differences in Young Children's Moral Judgments."
45. N. Chernyak and D. M. Sobel, "'But He Didn't Mean to Do It': Preschoolers Correct Punishments Imposed on Accidental Transgressors," *Cognitive Development* 39 (2016): 13–20.
46. Vaish, A., M. Missana, and M. Tomasello. "Three-Year-Old Children Intervene in Third-Party Moral Transgressions." *British Journal of Developmental Psychology* 29 (2011): 124–30.
47. D. M. Sobel and N. Z. Kirkham, "Bayes' Nets and Babies: Infants' Developing Statistical Reasoning and Their Representation of Causal Knowledge," *Developmental Science* 10 (2007): 298–306; D. M. Sobel and N. Z. Kirkham, "Blickets and Babies: The Development of Causal Reasoning in Toddlers and Infants," *Developmental Psychology* 42 (2006): 1103–15.
48. Wellman, *Making Minds*.
49. Y. Bar-Haim et al., "Nature and Nurture in Own-Race Face Processing," *Psychological Science* 17 (2006): 159–63; F. Warneken and M. Tomasello, "The Roots of Human Altruism," *British Journal of Psychology* 100 (2009): 455–71; H. Over, "The Influence of Group Membership on Young Children's Prosocial Behavior," *Current Opinion in Psychology* 20 (2018): 17–20.
50. T. F. Pettigrew and L. R. Tropp, "A Meta-analytic Test of Intergroup Contact Theory," *Journal of Personality and Social Psychology* 90 (2006): 751–83; N. Dasgupta and L. M. Rivera, "When Social Context Matters: The Influence of Long-Term Contact and Short-Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions," *Social Cognition* 26 (2008): 112–23.
51. R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The Brain's Default Network: Anatomy, Function, and Relevance to Disease," *New York Academy of Sciences* 1124 (2008): 1–38.
52. Ibid.; G. Sevinc, and R. N. Spreng, "Contextual and Perceptual Brain Processes Underlying Moral Cognition: A Quantitative Meta-analysis of Moral Reasoning and

Moral Emotions," *PLOS ONE* 9, no. 2 (2014): e87427, https://doi:10.1371/journal/pone.0087427.

53. R. L. E. P. Reniers et al., "Moral Decision-Making, ToM, Empathy, and the Default Mode Network," *Biological Psychiatry* 90 (2012): 202–10; W. Chiong et al., "The Salience Network Causally Influences Default Mode Network Activity during Moral Reasoning," *Brain* 136 (2013): 1929–41; B. Garrigan, A. L. R. Adlam, and P. E. Langton, "Neural Correlates of Moral Decision-Making: A Systematic Review and Meta-analysis of Moral Evaluations and Response Decision Judgments," *Brain and Cognition* 108 (2016): 88–97, corrigendum, *Brain and Cognition* 111 (2016): 104–6.
54. A. Rangell, C. Camerer, and P. R. Montague, "A Framework for Studying the Neurobiology of Value-Based Decision-Making," *Nature Reviews: Neuroscience* 9 (2008): 545–56; T. E. J. Behrens et al., "Associative Learning of Social Value," *Nature* 456 (2008): 245–50; A. Shenhav and J. D. Greene, "Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude," *Neuron* 67 (2010): 667–77; F. A. Cushman, and L. Young, "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations," *Cognitive Science* 35 (2011): 1052–75.
55. J. D. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science* 293 (2001): 2015–18; J. Greene and J. Haidt, "How (and Where) Does Moral Judgment Work?," *Trends in Cognitive Sciences* 6 (2002): 517–23; J. D. Greene et al., "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment," *Cognition* 111 (2009): 364–71.
56. L. Pessoa, "On the Relationship between Emotion and Cognition," *Nature Reviews Neuroscience* 9 (2008): 148–58.
57. S. R. Quartz, "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling," *Trends in Cognitive Sciences* 13 (2007): 209–15; A. D. Craig, "How Do You Feel—Now? The Anterior Insula and Human Awareness," *Nature Reviews Neuroscience* 10 (2009): 59–70; R. M. Nesse and P. E. Ellsworth, "Emotion, Evolution, and Emotional Disorders," *American Psychologist* 64 (2009): 129–39.
58. Cf. Marvin Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (New York: Simon and Schuster, 2006).
59. For some studies supporting the reliability of such electronic in-class sampling, see J. R. Stowell and J. M. Nelson, "Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion," *Teaching of Psychology* 34 (2007): 253–58; G. E. Kennedy and Q. I. Cutts, "The Association between Students' Use of an Electronic Voting System and Their Learning Outcomes," *Journal of Computer Assisted Learning* 21 (2005): 260–68.
60. J. J. Thomson, "Killing, Letting Die, and the Trolley Problem," *Monist* 59 (1976): 205–17.
61. N. Gold, A. M. Colman, and B. D. Pulford, "Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems," *Judgment and Decision Making* 9 (2014): 65–76; C. D. Navarette et al., "Virtual Morality: Emotion and Action in a Simulated 'Trolley Problem,'" *Emotion* 12 (2012): 364–70.

62. For an especially sophisticated discussion, see F. Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007).
63. Though on the potential influence of order effects on Loop verdicts, see S. M. Liao et al., "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning," *PNAS* 110, no. 11 (2013): E1064–E1073, https://doi/10.1073/pnas.1214869110.
64. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment"; F. A. Cushman, "Action, Outcome, and Value in a Dual-System Framework for Morality," *Personality and Social Psychology Review* 17 (2013): 273–92.
65. Greene, J.D., F.A. Cushman, L.E. Steward, K. Lowenberg, L.E. Nystrom, and J.D. Cohen, "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgments," *Cognition* 111 (2009): 364-71.
66. Bartells, D.M. and D.A. Pizarro, "The Mismeasure of Morals: Antisocial Personality Traits Predict Utilitarian Responses to Moral Dilemmas," *Cognition* 121 (2012): 490-503; Kahane, G., J. A. C. Everett, B. D. Earp, M. Farias, and J. Savulescu. "'Utilitarian' Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good." *Cognition* 134 (2015): 193–209.; Y. Gao and S. Tang, "Psychopathic Personality and Utilitarian Moral Judgment in College Students," *Journal of Criminal Justice* 41 (2013): 342–49; P. Conway and B. Gawronski, "Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach," *Journal of Personality and Social Psychology* 104 (2013): 216–35; E. Gleichgerrcht and L. Young, "Low Levels of Empathic Concern Predict Utilitarian Moral Judgment," *PLOS-One* 8 (2013): e60418. But see also for qualifications P. Conway, J Goldstein-Greenwood, D. Polacek, and J.D. Green, "Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers," *Cognition* 179 (2018): 241–65..
67. Aristotle, *Nicomachean Ethics* (350–340 BCE), trans. T. Irwin, 2nd ed. (Indianapolis, IN: Hackett, 1999).
68. David Hume, *An Enquiry concerning the Principles of Morals* (1751), ed. T. L. Beauchamp (Oxford: Oxford University Press, 1998); David Hume, *A Treatise of Human Nature* (1738), ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Oxford University Press, 1978).
69. Buckner, Andrews-Hanna, and Schacter, "The Brain's Default Network."
70. M. E. P. Seligman et al., "Navigating into the Future or Driven by the Past?," *Perspectives in Psychological Science* 8 (2013): 119–41.
71. Including a study of the "Knobe Effect": C. S. Sripada, "Mental State Attributions and the Side-Effect Effect," *Journal of Experimental Social Psychology* 48 (2012): 232–38.
72. E. L. Uhlmann, L. Zhu, and D. Tannenbaum, "When It Takes a Bad Person to Do the Right Thing," *Cognition* 126 (2013): 326–34.
73. For further discussion, see Peter Railton, "The Affective Dog and Its Rational Tale: Intuition and Attunement," *Ethics* 124 (2014): 813–59; Peter Railton, "Moral Learning: Conceptual Foundations and Normative Significance," *Cognition* 167 (2016): 172–90.
74. W. D. Ross, *The Right and the Good* (Oxford: Oxford University Press, 1930).

75. For an example in the confined world of games, see D. Silver et al., "A General Reinforcement Learning Algorithm Masters Chess, Shogi, and Go through Self-Play," *Science* 362 (2018): 1140–44.
76. Cf. Peterson et al., "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development."
77. The author would like to thank participants in the NYU Conference on Ethics and Artificial Intelligence (October 2016) for very helpful discussions, including especially Ned Block, Paul Boghossian, Nick Bostrum, David Chalmers, Vasant Dhar, Yann LeCun, S. Matthew Liao, Stuart Russell, Wendell Wallach, and Stephen Wolfram. I would also like to thank my colleagues Sarah Buss, Ben Kuipers, and Chandra Sripada for insightful conversations and sustaining encouragement.

References

Aristotle. *Nicomachean Ethics*. 350–340 BCE. Translated by T. Irwin. 2nd ed. Indianapolis, IN: Hackett, 1999.

Aslin, R. N., J. R. Saffran, and E. L. Newport. "Computation of Conditional Probability Statistics by 8-Month-Old Infants." *Psychological Science* 9 (1998): 321–24.

Axelrod, R., and D. Dion. "The Further Evolution of Cooperation." *Science* 242 (1988): 1385–90.

Bar-Heim, Y., T. Ziv, D. Lamy, and R. M. Hodes. "Nature and Nurture in Own-Race Face Processing." *Psychological Science* 17 (2006): 159–63.

Behne, T., M. Carpenter, J. Call, and M. Tomasello. "Unwilling versus Unable: Infants' Understanding of Intentional Action." *Development Psychology* 41 (2005): 328–37.

Behrens, T. E. J., L. T. Hunt, M. W. Woolrich, M. F. S. Rushworth. "Associative Learning of Social Value." *Nature* 456 (2008): 245–50.

Benhabib, S., and F. Dallmayr, eds. *The Communicative Ethics Controversy.* Cambridge, MA: MIT Press, 1990.

Bereczkei, T., P. Papp, P. Kinces, B. Bodrogi, G. Perlaki, G. Orsi, and A. Deak, "The Neural Basis of the Machiavellians' Decision Making in Fair and Unfair Situations," *Brain and Cognition* 98 (2015): 53-64.

Blair, R. J. R. "The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy." *Trends in Cognitive Sciences* 11 (2007): 387–92.

Blair, R. J. R. "The Emergence of Psychopathy: Implications for the Neurophysiological Approach to Developmental Disorders." *Cognition* 101 (2006): 414–42.

Boehm, C. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books,. 2012.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Buckner, R. L., J. R. Andrews-Hanna, and D. L. Schacter. "The Brain's Default Network: Anatomy, Function, and Relevance to Disease." *New York Academy of Sciences* 1124 (2008): 1–38.

Carpenter, M., J. Call, and M. Tomasello. "Twelve- and 18-Month-Olds Copy Actions in Terms of Goals." *Developmental Science* 8 (2005): F13–F20.

Chernyak, N., and D. M. Sobel. "'But He Didn't Mean to Do It': Preschoolers Correct Punishments Imposed on Accidental Transgressors." *Cognitive Development* 39 (2016): 13–20.

Chiong, W., S. M. Wilson, M. D'Esposito, A.S. Kayser, S.N. Grossman, P. Poorzand, W.W. Seeley, B.L. Miller, and K.P. Rankin. "The Salience Network Causally Influences Default Mode Network Activity during Moral Reasoning." *Brain* 136 (2013): 1929–41.

Conant, R. C., and W. R. Ashby. "Every Good Regulator of a System Must Be a Model of That System." *International Journal of Systems Science* 1 (1970): 89–97.

Conway, P., and B. Gawronski. "Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach." *Journal of Personality and Social Psychology* 104 (2013): 216–35.

Conway, P., J. Goldstein-Greenwood, D. Polacek, and J. D. Greene. "Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers." *Cognition* 179 (2018): 241–65.

Craig, A. D. "How Do You Feel—Now? The Anterior Insula and Human Awareness." *Nature Reviews Neuroscience* 10 (2009): 59–70.

Crockett, M., Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dayan. "Harm to Others Outweighs Harm to Self in Moral Decision Making." *PNAS* 111 (2014): 17320–25.

Cushman, F. A. "Action, Outcome, and Value in a Dual-System Framework for Morality." *Personality and Social Psychology Review* 17 (2013): 273–92.

Cushman, F. A., and L. Young. "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations." *Cognitive Science* 35 (2011): 1052–75.

Dasgupta, N., and L. M. Rivera. "When Social Context Matters: The Influence of Long-Term Contact and Short-Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions." *Social Cognition* 26 (2008): 112–23.

Decety, J., and A. N. Meltzoff. "Empathy, Imitation, and the Social Brain." In *Empathy: Philosophical and Psychological Perspectives*, edited by A. Copland and P. Goldie, 58–81. New York: Oxford University Press, 2011.

Devlin, J., M.-W. Chang, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, last revised May 24, 2019. arXiv:1810.04805v1. https://arxiv.org/abs/1810.04805.

Di Giorgio, E., J. L. Loveland, U. Mayer, O. Rosa-Salva, E. Versace, and G. Vallortigara. "Filial Responses as Predisposed and Learned Preferences: Early Attachment in Chicks and Babies." *Behavioural and Brain Research* 325 (2017): 90–104.

FeldmanHall, O., T. Dalgleish, D. Evans, and D. Mobbs. "Empathic Concern Drives Costly Altruism." *NeuroImage* 105 (2015): 347–56.

Frank, R. *Passions within Reason: The Strategic Role of the Emotions*. (New York: W.W. Norton, 1989).

Gao, Y., and S. Tang. "Psychopathic Personality and Utilitarian Moral Judgment in College Students." *Journal of Criminal Justice* 41 (2013): 342–49.

Garrigan, B., A. L. R. Adlam, and P. E. Langton. "Neural Correlates of Moral Decision-Making: A Systematic Review and Meta-analysis of Moral Evaluations and Response Decision Judgments." *Brain and Cognition* 108 (2016): 88–97; corrigendum, *Brain and Cognition* 111 (2016): 104–6.

Gleichgerrcht, E., and L. Young. "Low Levels of Empathic Concern Predict Utilitarian Moral Judgment." *PLOS-One* 8 (2013): e60418.

Gold, N., A. M. Colman, and B. D. Pulford. "Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems." *Judgment and Decision Making* 9 (2014): 65–76.

Goodman, N. D., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths. "A Rational Analysis of Rule-Based Concept Learning." *Cognitive Science* 32 (2008): 108–54.

Goodman, N.D., M.C. Frank, T.L. Griffiths, J.B. Tenenbaum, P.W. Battaglia, and J.B. Hamrick, "Relevant and Robust: A Response to Marcus and Davis," *Psychological Science* 26 (2015): 539-541.

Gopnik, A., and H. Wellman. "Reconstructing Constructivism: Causal Models, Bayesian Learning, and the Theory Theory." *Psychological Bulletin* 128 (2012): 1085–108.

Greene, J., and J. Haidt. "How (and Where) Does Moral Judgment Work?" *Trends in Cognitive Sciences* 6 (2002): 517–23.

Greene, J. D., F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen. "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment." *Cognition* 111 (2009): 364–71.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (2001): 2015–18.

Gupta, A. S., M. A. A. van der Meer, D. S. Touretzky, and A. D. Redish. "Hippocampal Replay Is Not a Simple Function of Experience." *Neuron* 65 (2010): 695–705.

Gweon, H., and L. Schulz. "16-Month-Olds Rationally Infer Causes of Failed Actions." *Science* 332 (2011): 1524.

Halberda, J. "The Development of a Word-Learning Strategy." *Cognition* 87 (2003): B23–B34.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. *The Foundations of Human Sociality: Economic Experiments and Ethnography Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press, 2004.

Hobbes, Thomas. *Leviathan*. 1651. Edited by C. B. MacPherson. London: Penguin, 1968.

Hoffman, M. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge: Cambridge University Press, 2001.

Hume, David. *An Enquiry concerning the Principles of Morals*. 1751. Edited by T. L. Beauchamp. Oxford: Oxford University Press, 1998.

Hume, David. *A Treatise of Human Nature*. 1738. Edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Oxford University Press, 1978.

Johnson, A., M. A. A. van der Meer, and A. D. Redish. "Integrating Hippocampus and Striatum in Decision-Making." *Current Opinion in Neurobiology* 17 (2007): 692–97.

Kahane, G., J. A. C. Everett, B. D. Earp, M. Farias, and J. Savulescu. "'Utilitarian' Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good." *Cognition* 134 (2015): 193–209.

Kavish, N., C. Bailey, C. Sharp, and A. Venta, "On the Relation between General Intelligence and Psychopathic Traits: An Examination of Inpatient Adolescents," *Child Psychiatry and Human Development* 49 (2018): 341-51.

Kamm, F. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.

Kennedy, G. E., and Q. I. Cutts. "The Association between Students' Use of an Electronic Voting System and Their Learning Outcomes." *Journal of Computer Assisted Learning* 21 (2005): 260–68.

Kidd, C., S. T. Piantadosi, and R. N. Aslin. "The Goldilocks Effect: Human Infants Allocate Attention to Sequences That Are Neither Too Simple nor Too Complex." *PLOS-One* 7 (2012): e36399.

Koenig, M. A., V. Tiberius, and K. Hamlin. "Children's Judgments of Epistemic and Moral Agents: From Situations to Intentions." Unpublished manuscript (2019).

Kwon, O.-S., and D. C. Knill. "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning." *PNAS* 110, no. 11 (2013): E1064–E1073. https://doi/10.1073/pnas.1214869110.

Liao, S. M., A. Wiegmann, J. Alexander, and G. Vong. "Putting the Trolley in Order: Experimental Philosophy and the Loop Case." *Philosophical Psychology* 25 (2012): 661–71.

Meltzoff, A. N. "'Like Me': A Foundation for Social Cognition." *Developmental Science* 10 (2007): 126–34.

Meltzoff, A. N., P. K. Kuhl, J. Movellan, and T. J. Sejnowski. "Foundations for a New Science of Learning." *Science* 325 (2009): 284–88.

Minsky, Marvin. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon and Schuster, 2006.

Moll, J., and M. Tomasello. "12- and 18-Month-Olds Follow Gaze to Spaces behind Barriers." *Developmental Science* 7 (2004): F1–F9.

Monagan, C., H. Bizumic, and M. Sellbom, "Nomological Network of Two-Dimensional Machiavellianism," *Personality and Individual Differences* 130 (2018): 161-72.

Moser, E. I., E. Kropff, and M.-B. Moser. "Place Cells, Grid Cells, and the Brain's Spatial Representation System." *Annual Review of Neuroscience* 31 (2008): 69–89.

Navarrette, C. D., M. M. McDonald, M. L. Mott, and B. Asher. "Virtual Morality: Emotion and Action in a Simulated 'Trolley Problem.'" *Emotion* 12 (2012): 364–70.

Nesse, R. M., and P. E. Ellsworth. "Emotion, Evolution, and Emotional Disorders." *American Psychologist* 64 (2009): 129–39.

Oderbank, S.G., J. Nitschke, A. Mokros, E. Habermeyer, and O. Wilhelm, "Psychopathic Men: Deficits in General Mental Ability, Not Emotion Perception," *Journal of Abnormal Psychology* 127 (2018): 294-304.

Over, H. "The Influence of Group Membership on Young Children's Prosocial Behavior." *Current Opinion in Psychology* 20 (2018): 17–20.

Pessoa, L. "On the Relationship between Emotion and Cognition." *Nature Reviews Neuroscience* 9 (2008): 148–58.

Peterson, C., V. Slaughter, C. Moore, and H. M. Wellman. "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development." *Developmental Psychology* 52 (2016): 46–57.

Pettigrew, T. F., and L. R. Tropp. "A Meta-analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology* 90 (2006): 751–83.

Pinker, S. *The Language Instinct*. New York: Morrow, 1994.

Pullum, G. K., and B. C. Scholz. "Empirical Assessment of Stimulus Poverty Arguments." *Linguistic Review* 19 (2002): 9–50.

Quartz, S. R. "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling." *Trends in Cognitive Sciences* 13 (2007): 209–15.

Railton, Peter. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics* 124 (2014): 813–59.

Railton, Peter. "Moral Learning: Conceptual Foundations and Normative Significance." *Cognition* 167 (2016): 172–90.

Rand, D. G., J. D. Greene, and M. A. Nowak. "Spontaneous Giving and Calculated Greed." *Nature* 489 (2012): 427–30.

Rangell, A., C. Camerer, and P. R. Montague. "A Framework for Studying the Neurobiology of Value-Based Decision-Making." *Nature Reviews: Neuroscience* 9 (2008): 545–56.

Reniers, R. L. E. P., R. Corcoran, B. A. Vollm, A. Mashru, R. Howard, and P. F. Liddle. "Moral Decision-Making, ToM, Empathy, and the Default Mode Network." *Biological Psychiatry* 90 (2012): 202–10.

Redish, A. D. "Vicarious Trial and Error." *Nature Reviews: Neuroscience* 17 (2016): 147–59.

Rilling, J. K., D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts. "A Neural Basis for Social Cooperation." *Neuron* 36 (2002): 395–406.

Ross, W. D. *The Right and the Good*. Oxford: Oxford University Press, 1930.

Roth-Hanania, R., M. Davidov, and C. Zhan-Waxler. "Empathy Development from 8 to 16 Months: Early Signs of Concern for Others." *Infant Behavior and Development* 34 (2011): 447–58.

Saby, J. N., A. N. Meltzoff, and P. J. Marshall. "Infant's Somatotopic Neural Responses to Seeing Human Actions: I've Got You under My Skin." *PLOS ONE* 8, no. 10 (2013): e77905. https://doi:10.1371/journal.pone.0077905.

Seligman, M. E. P., P. Railton, R. Baumeister, and C. S. Sripada. "Navigating into the Future or Driven by the Past?" *Perspectives in Psychological Science* 8 (2013): 119–41.

Sevinc, G., and R. N. Spreng. "Contextual and Perceptual Brain Processes Underlying Moral Cognition: A Quantitative Meta-analysis of Moral Reasoning and Moral Emotions." *PLOS ONE* 9, no. 2 (2014): e87427. https://doi:10.1371/journal/pone.0087427.

Shenhav, A., and J. D. Greene. "Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude." *Neuron* 67 (2010): 667–77.

Silver, D., T. Hubert, J. Schrittweiser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. "A General Reinforcement Learning Algorithm Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (2018): 1140–44.

Smetana, J. G., W. M. Rote, M. Jambon, M. Tasopoulos-Chan, M. Villalobos, and J. Comer. "Developmental Changes and Differences in Young Children's Moral Judgments." *Child Development* 83 (2012): 683–96.

Sobel, D. M., and N. Z. Kirkham. "Bayes' Nets and Babies: Infants' Developing Statistical Reasoning and Their Representation of Causal Knowledge." *Developmental Science* 10 (2007): 298–306.

Sobel, D. M., and N. Z. Kirkham. "Blickets and Babies: The Development of Causal Reasoning in Toddlers and Infants." *Developmental Psychology* 42 (2006): 1103–15.

Spelke, E. S., and K. D. Kinzler. "Innateness, Learning, and Rationality." *Child Development Perspectives* 3 (2009): 96–98.

Sripada, C. S. "Mental State Attributions and the Side-Effect Effect." *Journal of Experimental Social Psychology* 48 (2012): 232–38.

Stowell, J. R., and J. M. Nelson. "Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion." *Teaching of Psychology* 34 (2007): 253–58.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (2011): 1279–85.

Thomson, J. J. "Killing, Letting Die, and the Trolley Problem." *Monist* 59 (1976): 205–17.

Todorov, E. "Optimality Principles in Sensorimotor Control." *Nature Neuroscience* 7 (2004): 907–15.

Todorov, E., and Z. Ghahramani. "Unsupervised Learning of Sensory-Motor Primitives." *Proceedings of the 25th Annual International Conference of the IEEE EMBS* (2003): 1750–53.

Turiel, E. *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press, 2002.

Uhlmann, E. L., L. Zhu, and D. Tannenbaum. "When It Takes a Bad Person to Do the Right Thing." *Cognition* 126 (2013): 326–34.

Vaish, A., M. Missana, and M. Tomasello. "Three-Year-Old Children Intervene in Third-Party Moral Transgressions." *British Journal of Developmental Psychology* 29 (2011): 124–30.

Van den Oord, A., Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding." *arXiv*, last revised January 22, 2019. arXiv:1807.03748v1.

Warneken, F., and M. Tomasello. "Altruistic Helping in Human Infants and Young Chimpanzees." *Science* 311 (2006): 1301–3.

Warneken, F., and M. Tomasello. "The Roots of Human Altruism." *British Journal of Psychology* 100 (2009): 455–71.

Wellman, H. M. *Making Minds: How Theory of Mind Develops*. Oxford: Oxford University Press, 2014.

Wellman, H. M., and D. Liu. "Scaling of Theory-of-Mind Tasks." *Child Development* 75 (2004): 523–41.

2

The Use and Abuse of the Trolley Problem

Self-Driving Cars, Medical Treatments, and the Distribution of Harm

F. M. Kamm

In this chapter I first briefly present cases that are standardly considered "Trolley Problem" Cases along with standard moral judgments about permissible conduct in these cases. Next, I consider the ways in which many standard car driving cases differ as a conceptual matter from standard Trolley Problem Cases with which some compare them. I argue that the cases involving cars raise distinctive moral issues different from the distinctive issues raised by standard Trolley Problem Cases. I also consider how some medical cases differ from some standard trolley cases with which some compare them. Finally, I discuss some moral issues raised by self-driving cars by comparison to Trolley Problem Cases, including the role of those who would program the cars and the liability to harm of pedestrians, drivers, and passengers.

2.1. A Hypothetical Case

I once considered what I called the Ambulance Case.[1] In it society was to decide ex ante (i.e., in advance of knowing who would be affected one way or another) how an ambulance should be programmed when it came to a choice between saving people by rushing them to the hospital and harming pedestrians on the route or letting the patients die but harming no pedestrians. I imagined that the ambulance could be made to detect how many people it was carrying and how many pedestrians would be harmed, and, to simplify matters, I assumed (as I do here) that the life of each person was at stake and that they were alike in morally relevant respects. One question I considered was whether we should deliberately program our ambulance carrying five people to continue on its route by having the program disable a stopping mechanism whenever more lives would be lost by its stopping than by running over one person on the road. I argued that even though ex ante (at the time we decide on how to program the ambulance) each person in the society would maximize his chances of survival by not allowing the

F.M. Kamm, *The Use and Abuse of the Trolley Problem* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). © Oxford University Press. DOI: 10.1093/oso/9780190905040.003.0003

ambulance to stop, it could be wrong to program the ambulance in this way, just as it would be wrong for a driver in control of the ambulance to drive over the one person to get the five to the hospital. Certainly, the fact that only a program, not a person, would disable the stopping mechanism at the time of impact does not remove responsibility for this happening from the people who programmed the ambulance.

2.2. Standard Trolley Cases

Recently, reality has caught up with hypothetical cases like the Ambulance Case in which vehicles can be programmed to move in various ways. One such instance is the design of so-called self-driving cars of which there could be at least two types: (1) those that have no person at all driving them and that operate completely on the program designed for them (call this the Complete Case) and (2) those that have a person driving them but whose program can override or supplement a driver's control at crucial points (e.g., the car will stop despite the driver trying to continue; call this the Partial Case).[2]

Some have thought that what is known as the Trolley Problem, a topic in normative ethical theory, might help us with the practical problem of creating programs for self-driving cars. Here is a description of the basic Trolley Case created by Philippa Foot:[3] A driver is on his out-of-control trolley that is headed toward killing five innocent people on a track. To save them he can only turn the trolley to a sidetrack where, he foresees, the trolley will instead kill one other person standing on that track. A variation on this case introduced by Judith Thomson involves a bystander, not the driver, deciding whether to turn the trolley (with the same effect) when the driver cannot do so (the Bystander Case).[4] It is commonly thought that since the driver started the trolley, for him it is a choice of killing five or killing one. However it might be argued that if, independently of any act or omission of his, the trolley goes out of control, how he comes to cause the death of the five would be very different from his deliberately turning the trolley when he foresees it will kill someone else. For the bystander it is a choice of letting five die or killing one.[5] In none of these cases is the person who turns the trolley or who is on the trolley at risk of dying. In none of these cases is the person who would be killed by the redirected trolley the cause of its threatening the five or obligated to either give his life or assume a position in which he will be killed as a side effect in order to save the five other people. He would also lose his life by someone else imposing the loss on him, not by his imposing it on himself.

Many have thought that in these cases the driver is morally required and the bystander is at least permitted to do what will minimize the number of people

who would be killed by redirecting the trolley and that there is some principle that explains the obligation and permissibility, respectively. (These cases involve a choice between five or one being killed, but the argument that would justify killing one to save five arguably would also justify redirecting toward one person to save even two other people and redirecting toward any smaller number of people to save a larger number of other people. Indeed I think that if an argument could not justify killing one to save two from the trolley, it could not justify killing one to save five from it.)

By contrast, many think it would be impermissible for either the driver or the bystander to do what is necessary to save the five people by killing one other person in the following ways: (1) topple someone from a bridge in front of the trolley so that his body stops it and the trolley kills him (Topple Case)[6] or (2) use a small bomb whose explosion would stop the trolley from hitting the five, though a piece of the bomb would fly off, killing an innocent pedestrian (Bomb Case). It is thought that some principle explains why these things are impermissible even though they would also minimize the number of innocent people killed. We could think of the Trolley Problem as explaining why killing is permissible in the Trolley and Bystander Cases but not in the Topple and Bomb Cases. This problem is part of the more general debate between act consequentialists, who think it always permissible to bring about the best consequences, and nonconsequentialists, who deny this.

2.3. Other Cases

I have described what philosophers who have discussed the Trolley Problem take to be some standard cases involved in it. Let us now consider how these cases compare conceptually with those that have been described as Trolley Problem Cases by others, including contributors to popular internet sites who aim to acquaint the general public with the problem in connection with ordinary and self-driving cars. At this stage, I am concerned only with conceptual similarities and differences and put to one side moral judgments of the cases. Nevertheless I am concerned with similarities and differences because they may be morally significant. For one example, Chris Rampolla (who describes himself as "a philosopher by training") says the following in his online article:

> Suppose that you're driving your car in the right hand lane of a one-way street on a winter evening. As you approach a red light at an intersection, you tap the brakes and begin to skid. Ahead of you the left lane is closed and is blocked by a concrete barrier in front of a crosswalk. There are no obstructions in the right lane. A pedestrian has legally entered the crosswalk on the right side of

> the street and is attempting to cross over to the other side. You have just enough time and just enough control of the car to make a decision about which lane to enter but you cannot stop your car. Should you choose to continue on in the right lane, the pedestrian will be struck by the car and will likely die. Should you choose to direct your car into the left lane, the collision of your car with the barrier will save the life of the pedestrian but will very likely kill you, the driver. What do you choose to do? Now ask yourself that same question, except this time consider that your child is in the car and would likely die from an impact with the barrier. Next consider that the pedestrian in the road is also accompanied by a child. Still further, consider that this time your spouse and child are in your car with you and there are three elderly people in the crosswalk. Has your choice changed? More generally, what is the moral thing to do in each of these situations and is there any commonality between them? These are modern versions of a philosophical problem known as *The Trolley Problem*.[7]

2.3.1. Threats and Nonthreats

Rampolla here describes the Trolley Problem as being instantiated in a case where a driver of a car that cannot be stopped has to decide whether to let the car continue on to (very likely) kill a pedestrian, or turn it and (very likely) kill himself.[8] (Henceforth, to simplify, I will assume someone will definitely be killed, as is assumed in the standard trolley cases.)[9] But the standard Trolley Problem Cases discussed by philosophers are about whether one may kill some innocent, nonthreatening person so as to either not kill or save other innocent, nonthreatening people from a threat already facing them. These cases do not, as in Rampolla's case, involve someone who is presenting a threat of death to others (the driver) deciding whether to kill innocent, nonthreatening people rather than kill himself. Philosophers might refer to the latter sort of case as an "innocent threat case." Such cases are distinguished by the threatener himself, a bystander, or a potential victim having to decide whether to harm the threatener, who became such despite being innocent of any moral wrongdoing, or allow him to proceed to harm some nonthreatening people.

I am not concerned with legislating the use of the terms "Trolley Problem" and "Innocent Threat Case." I am concerned with the possibility that the difference I have identified between standard trolley cases and the one in Rampolla's example may be morally significant and that it is wrong to assume without additional argument that a driver who threatens people should weigh his life or have it weighed by others as the lives of innocent, nonthreatening persons should be weighed when one decides who will be harmed. To keep this moral issue in mind, I shall distinguish the Trolley and Innocent Threat Cases in the ways I have just described (though others may follow a different practice).[10]

There are different types of Innocent Threat Cases which make it clearer why the threatener is considered morally innocent. Robert Nozick described a man unwillingly shot out of a cannon, headed toward landing safely on someone who, however, would be killed by the impact.[11] May the potential victim or a bystander kill the innocent threat to prevent his fall from killing the victim? In other Innocent Threat Cases (sometimes called cases of minimal responsibility for harm)[12] the threatener is not totally inactive in causing the threat but is not at fault in his actions (e.g., a psychotic killer, a nonnegligent driver whose brakes fail in his well-maintained car). All these cases involve unjustified threats, but might there be other Innocent Threat Cases where a nondeliberate threat is justified? For example, consider a pilot fighting on the just side in a war who unintentionally bombs a military facility that it would have been permissible to intentionally bomb given that there is only permissible collateral harm to civilians. In standard just war theory anyone who presents a threat is considered a "noninnocent," but morally speaking one might consider him innocent. May these threatened civilians nevertheless shoot him down to protect themselves (especially if this will not affect the good military effect of his bombing)? In this case it is the pilot who is the threat, for if we could eliminate him and thus his control of the bombs, the threat to civilians would stop. However, in cases involving an already out-of-control car it is strictly the vehicle that is the threat; getting rid of the driver may not stop the threat,[13] though doing something to the vehicle that as a side effect will kill the driver will stop the threat. Nevertheless I think in an extended sense, the driver in such cases is an innocent threat if he was driving the vehicle before he lost control of it.

In standard Innocent Threat Cases, the choice is between the original potential victim(s) being harmed and the threatener being harmed (whether the harm to him would be imposed by the threatener himself or by someone else). Suppose a third option is added so that the threat can be redirected to another nonthreatening victim. Then we could possibly have a combined Innocent Threat and Trolley Problem Case as I am using these terms. However, if a driver had to bear the burden of being harmed rather than his original potential victims, it is unlikely to be permissible for anyone except the new potential victim to decide that the threat should be turned to her, holding losses to all constant. Then problems distinctive to standard Trolley Problem Cases would not arise.

2.3.2. Imposing Harm versus Having It Imposed

Note also that in the standard Trolley Cases, unlike the case I cite from Rampolla, the person who decides what to do with the trolley (either the driver or a bystander) is not one of those who might be harmed. Hence those who stand to be

harmed in standard Trolley Cases have harm imposed on them by others rather than imposing it on themselves. By contrast, in Rampolla's case the driver faces the option of harming himself.[14] One should not assume without argument that what may be imposed on one is the same as what one has a duty to impose on oneself.

Hence, unlike the standard Trolley Cases, Rampolla's case involves both the possibility of harm to the threatening driver and the driver having to decide whether to impose the harm on himself. The first factor may make it easier to justify harm to one person; the second factor may make it harder to demand it. Suppose one could show that the driver was permitted to treat himself like any nonthreatening possible victim. The fact that no such victim would have to redirect the threat away from others to himself would imply that the driver (who is actually the threat) would not have to redirect in a way that kills himself.[15]

2.3.3. Passengers and Varied Characteristics

As the quote from Rampolla shows, he also considers cases in which there are people in the threatening vehicle who are simply passengers and who stand to be harmed if the threatening driver himself decides to avoid harming pedestrians. Cases with passengers are unlike the standard Trolley Cases in which there is no one on the trolley who can be harmed, and it is also a complication of the standard Innocent Threat Case. If there are passengers in a threatening vehicle, it would be another mistake to assume without argument that they are to be treated as morally equivalent to pedestrians who would be hit by the threat and that their lives should be weighed in the same way as pedestrians.

Some of Rampolla's cases also assign different qualities to the different people who might be harmed (e.g., some are old, some young, some are strangers, some are one's spouse or child).[16] By contrast the standard Trolley and Innocent Threat Cases abstract from such distinctions in order to determine (1) whether in the Trolley Cases differences in how we come to kill people (e.g., by redirection, by toppling, etc.) make a moral difference and (2) whether in the Innocent Threat Cases the difference between being a threat (albeit morally innocent) rather than a pedestrian makes one more liable to be harmed. As in science, it has been thought in philosophy that to test for the moral significance of a factor, we should insofar as possible hold other variables constant. Nevertheless, varying additional characteristics to see how they affect one's decisions about sparer cases need not be a methodological mistake since otherwise crucial factors can be overridden or have their effect altered in changing contexts.

2.3.4. Possible Principles

A final point about Rampolla's discussion is that he describes Foot as supporting a revised version of the Doctrine of Double Effect (DDE) to justify the driver turning the trolley. He says:

> A modern interpretation of the doctrine of double effect was put forth by Philippa Foot in 1967. The problem (at the time) had nothing to do with driving, but instead was one of a number of thought experiments she used to examine the morality of abortion. As an example of *double effect* [emphasis added] she suggested the following:
>
> "The steering driver faces a conflict of negative duties since it is his duty to avoid injuring five men and also his duty to avoid injuring one. In the circumstances he is not able to avoid both, and it seems clear that he should do the least injury he can. The judge, however, is weighing the duty of not inflicting injury against the duty of bringing aid. He wants to rescue the innocent people threatened with death but can do so only by inflicting injury himself. Since one does not in general have the same duty to help people as to refrain from injuring them, it is not possible to argue to a conclusion about what he should do from the steering driver case."

But it is not true that the distinction Foot draws in the part of her article quoted by Rampolla is a modern interpretation of the DDE. The DDE distinguishes morally between harm that happens as a side effect and harm that is intended. It claims the latter is impermissible even as a means to a greater good while the former at least does not rule out pursuing a greater good. Foot proposes an alternative to the DDE that has nothing to do with whether one intends harm (or harm is a means). It focuses on a moral distinction between harming and not aiding even when harm is merely foreseen as a side effect and neither intended nor causally a means. For example, Foot thought that using a gas in surgery to save five people from their illness is ruled out if this gas will also cause someone else's death as a mere foreseen side effect. (I will call this the Gas Case.) (Note that her claim that the duty not to harm is stronger than the duty to aid also might rule out turning the trolley in the Bystander Case since the bystander would harm one person to aid five.)[17]

2.4. New and Old Threats?

Other examples of the problematic use of the Trolley Problem occur in medical ethics discussions. One instance is in the work of Dr. Marya Zilberberg.[18]

She correctly identifies a version of the Trolley Problem as reconciling the apparent permissibility of the driver diverting the trolley and the impermissibility of a bystander toppling a person in front of the trolley to stop it even though in both cases five people would be saved from death and one person would be killed. She goes on to compare these cases with the use of mammography which is said to save eight women who would otherwise have died from cancer for every thousand mammogrammed but lead to the death of at least one woman who would not otherwise have died (due to false positives). She says about this Mammography Case:

> Well, then we have the trolley problem, don't we? We are potentially sacrificing 1 individual to save 8. And who does the sacrificing is where the variations of the trolley problem come in. . . . The payer certainly sees this issue as the original formulation of the problem: Why not throw this financial switch to achieve net life savings? But for a clinician who deals with the individual patient this may be akin to pushing her over the bridge toward a potentially fatal event.

The first thing to note is that unlike the driver in the "original formulation of the problem," the payer did not have a role in causing the threat that must be dealt with (in this case, cancer). Nor is he diverting the cancer that threatens the life of some women toward fewer women. The payer is helping to pay for a new means to help some women (the mammogram) which presents a bad side effect threat to a smaller number of other women. (This is analogous to using the small bomb to stop the trolley when the bomb would kill another person.) This also implies that the doctor who orders mammograms that as a side effect harm a woman is not harming a woman as a *means* to help save eight others, which would be akin to toppling someone in front of the trolley as a means of stopping it. This is shown by the fact that if mammograms the doctor orders did not cause the harm to one woman as a side effect, this would not reduce mammograms' effectiveness in saving eight out of a thousand women; the threat and harm to the one is not needed to save the eight.

An additional aspect of the Mammography Case to which Dr. Zilberberg points is that ex ante each person could be either one of the eight who will be benefited or the one who will be harmed. Hence she thinks it is important to understand the patient's attitude to risk. And indeed introducing a means that will unavoidably risk harm as a side effect to each of those who, as far as we know ex ante, also get from that means a greater chance to benefit may be permissible. However, this does not imply that it is permissible to use means that will help others when it is known that the means will as a side effect impose certain death on a particular other person, and it is still possible to prevent its doing so by not using the means at the time we know it would present the threat (as in

the Bomb and Gas Cases). This is so even if it is permissible in the Trolley and Bystander Cases to impose certain death on a particular other person when one could avoid it (independent of any ex ante calculation of risks and benefits). As already noted, it is one aspect of the Trolley Problem to explain these differences in permissibility.

Not distinguishing between the Trolley Case and cases like Bomb and Gas is also exemplified by some discussions in medical ethics that try to analogize the use of electronic cigarettes to diverting the trolley.[19] That is, many people will die of smoking. Suppose we can reduce their numbers by converting them to use of electronic cigarettes, which are somewhat less bad for them. However, suppose also that as a side effect of this policy a smaller number of other adults who would not have smoked will also take up electronic cigarettes, thus becoming worse off than they would otherwise have been. (This is a hypothetical case insofar as it abstracts from real effects, especially on underage users.) This case is not analogous to diverting the trolley away from more people to fewer people for the same reason that the Bomb and Gas Cases are not like diverting the trolley. In this case there is a prior threat (of cigarettes) to many people. If we introduced e-cigarettes as a means to help them, we would not be diverting cigarette use but introducing a new means that would have the side effect of harming (by hypothesis) a smaller number of people who never smoked. We may wonder whether this conceptual difference makes a moral difference, but it *is* a conceptual difference. It is just important to remember that not all cases that involve only foreseeably killing fewer nonthreatening people to save a greater number of other nonthreatening people are like the standard Trolley Case in which diverting a threat seems permissible.

However, it is worth noting that, as described, the Electronic Cigarette Case differs from the Gas and Bomb Cases in at least two significant ways: (1) The e-cigarettes that reduce deaths among cigarette smokers do not directly cause harm to others, as do the bomb and the gas. It is only because an intervening agent takes up smoking e-cigarettes that he may be harmed. Helping the cigarette smokers by getting them to use e-cigarettes at most *enables* the harm to another new group of e-smokers by making available to them the new option of e-cigarettes. (2) The likelihood of death to each of the new e-cigarette smokers is (assumed) less than the likelihood of death to the original cigarette smokers. (This is so even if there are eventually more deaths due to e-cigarettes than to cigarettes because more rather than the hypothesized fewer new smokers will use them.) So unlike the Bomb and Gas Cases, the probability of death occurring to each of the newly threatened people is less than the probability of death occurring to each of the people originally threatened by cigarettes. These two factors may make the grounds that rule out using the bomb and gas inadequate to rule out the introduction of e-cigarettes in the hypothesized circumstances despite their bad side effects.

Here is an implication for self-driving cars of what we have just said about the Mammography and Electronic Cigarette Cases: The permissibility of killing some to save others in the standard Trolley Case is relevant to programming cars only if programming is about redirecting the threat that the car itself presents. However, this does not mean that to be morally like the Trolley Case anyone who dies as a result of a threatening car being redirected must also be killed by that very car. For example, suppose that the turning trolley (or car) caused a new threat of a rockslide that killed a pedestrian. Turning the trolley (or car) is still a permissible solution in the trolley-type case because it is the trolley (or car) turning away from the five people that causes the new threat,. This is not so with a newly introduced means to turn the trolley like the bomb which presents a new threat.[20]

The overall conclusion of considering these discussions that try to make use of the Trolley Problem (in both sections 2.2 and 2.3) is that they often fail to recognize the very distinctions the Trolley Problem Cases are about. These distinctions may or may not be morally significant, but not recognizing them is itself problematic.

2.5. Particular Moral Issues in Self-Driving Cars

Though some would disagree, let's assume for the sake of argument that all the judgments I cited in section 2.2 about what is commonly thought to be permissible, obligatory, and prohibited in the standard Trolley Problem Cases are correct and that there are principles that justify these judgments.[21] This would imply that we know what morally should or may be done in many cases. If we do not have access to the principles that underlie our judgments, we cannot program the principles into cars, though we might provide rules for what they should do in a variety of cases.[22] However, there may be other ways in which cars can "learn" what to do besides being programmed with rules or principles. It is said that, like people, machines exposed to various situations can self-learn to make correct choices without following explicit rules or interpreting known principles. These are complex issues about learning that I will not discuss here. For simplicity I will refer to the cars in Complete and Partial Cases as being programmed with principles that lead them to behave properly.[23] My concern here is not *how* to make self-driving cars behave properly but *what* the proper way for them to behave is and whether there are substantive moral (rather than merely conceptual) differences between what is morally proper in standard Trolley Cases and in the Complete and Partial Cases.

Here are some issues to consider.[24]

2.5.1. Why Have Self-Driving Cars?

The primary benefit of Complete or Partial cars is that they will prevent deaths by preventing situations in which any person's life is threatened. Their primary benefit is not to merely reduce lives lost once something has already gone wrong and at least someone will have to die.

In the Trolley Cases something has already gone wrong since the trolley has gone out of the driver's control with respect to the five, and the question is who is to die when someone must die. Hence the primary issue with which the Trolley Problem (and also Innocent Threat Cases) deal—what to do when someone must die—would arise only when self-driving cars have failed to satisfy the primary reason for having them. If self-driving cars got into dangerous situations more often than cars completely under a driver's control, the fact that they could be programmed to do a better job than a driver of minimizing the harm they would cause would not speak as strongly in their favor. However, if these cars got into life-threatening situations less often than human-driven cars, that they were worse in determining who will die when someone must might not speak very strongly against them.[25]

2.5.2. Do Moral Principles That Apply to Persons Apply to Machines?

It may be said by some that the sorts of moral prohibitions that make it impermissible, for example, to kill one person to save five in the Topple and Bomb Cases, are relevant to the conduct of persons but not to machines because the prohibitions are grounded in "the agent's personal point of view" of acting in certain ways and automated cars are not agents that have a personal point of view. This so-called agent-relative view of the ground of prohibition on people harming people has been defended by Thomas Nagel in some of his work.[26] It might also be said that there can be *reasons* to do one thing rather than another only for persons because they can have conscious appreciation of considerations for and against acting in some way, but there are no such "reasons for" machines if they have no conscious appreciation of considerations for and against acting in some way. Another ground for thinking that moral principles that apply to people do not apply to machines is that human persons would have emotional responses to and struggle emotionally with killing in some ways, and this cost to them should be taken into account in determining what they should do. But machines do not have emotional responses or struggles that should be taken into account in determining what they should do.

Consider objections to these three views in turn. Suppose that the ground of the wrongness of acting in Topple and Bomb is not concern for the agent's personal view of acting in certain ways (or even for the relationships between agent and victim generated by acting in certain ways). Rather suppose the wrongness is grounded in concern for the potential victim's status as a being who may not be treated in certain ways.[27] Then the potential victim's status could be violated as much by a machine as by a person. It is interesting and important that which theory correctly grounds nonconsequentialist prohibitions on harming could be relevant to which principles to use in programming machines. For purposes of this discussion I shall assume that the "victim-focused" account of prohibitions is correct.

Second, reasons for an entity's behaving in a certain way can exist independently of an entity's awareness of this. Certainly there can be a reason for a person not to drink a poisoned liquid though he isn't aware it is poisoned. Could there similarly be a reason for a machine not to kill in Topple if a person's status provides grounds for his not being killed in that way? We could at least say that there is a reason for there not to be entities that would kill someone in this way and people who were aware of these reasons might be morally required to interfere with the machine that would kill in this way. This could be true even if the machine was not designed by people but fell like manna from heaven or grew like a plant.

The fact that machines would not be affected emotionally by their harming others is irrelevant to the permissibility of their movements if the impermissibility of people behaving in comparable ways has nothing to do with the emotional costs to them of doing so. This would be so if the reason for not behaving in that way stems from the status of the potential victim and the emotional costs arise from implicit recognition by agents that they have acted impermissibly. If people could take a pill that made them not react emotionally to their killing in Topple, this would not affect the impermissibility of their so acting. (The absence of emotional effects, however, might make it easier for machines than people to do the right thing when the cost to people of doing so would be great, e.g., their own destruction. Machines would not have excuses that people might have for not doing the right thing when this involves damage or harm to them. I shall return to this point later.)

Finally, at least some self-driving cars would be programmed in advance to deal with any upcoming situation while a person driving a car would decide what to do when in the situation. Does this make what the car should do different from what a person should do in the same situation? In the Mammography Case it was said that an ex ante decision to use a diagnostic test in a population could be morally permissible even though we know that it will unavoidably harm someone at a time when we will be unable to help her. But that does not mean that we should

use such a mammogram test instead of one that could detect and interfere with harm it was about to cause. Similarly, that a car will be programmed in advance (unlike a human person) does not mean that it should not be programmed to behave in that situation to avoid harming someone in the way a person should. We need not program a device in advance to topple someone in the Topple Case because minimizing deaths would be ex ante in the interests of all if it would be wrong on victim-focused grounds for a person to commit in advance to doing the same thing at a time when she could still avoid doing it.[28]

2.5.3. Programmers and Company Duties

2.5.3.1. Programmers as Bystanders?

Do those who program cars they will not drive occupy a role analogous to that of the bystander in the Bystander Case? One difference between these two agents is that the bystander is dealing with a trolley that is already doubly out of the driver's control: the driver cannot control it as it heads to the five, and he lacks the ability to turn it away from the five to the sidetrack. By contrast, the programmer, at least in the Partial Case, is deciding whether to make the car be to some degree beyond a conscious driver's control. The program is designed not merely to recommend a course of action to the driver but to actually compel the car to make certain movements. (In deciding whether to make completely driverless cars, programmers are deciding whether there is to be any person driving at all.)

Prima facie, the Partial Case seems to raise special moral issues not raised by the Complete Case since it involves deliberately limiting the liberty of conscious agents to decide for themselves whether and how to prevent harm they would cause to others. The bystander in the following revised Bystander Case seems more like the programmer in a Partial Case: A bystander sees a trolley that can be stopped from killing five people only if it is turned where it will kill one other person. The driver retains the power to redirect the trolley, but she may or may not actually do this. The bystander presses a switch that takes the power to turn out of the driver's hands and puts it in his own hands. Call this the Intrusive Bystander Case. It raises at least two questions: (1) Is it permissible for a private person like the bystander to transfer power to himself? (2) Should a bystander as willingly take it upon himself to redirect the trolley, thereby killing a person, when it is possible that the driver herself would fulfill her responsibility to redirect if she retained power? (As we shall see, the answer to questions analogous to these is complicated in Partial Cases by the fact that, unlike what is true in the Trolley and Bystander Cases, the driver herself might be harmed as a result of a decision. Until further notice, I will assume that the driven cannot be harmed.)

The program that takes over in the Partial Case and those who create the program seem like intrusive bystanders. They become less intrusive if the driver has the option of turning the intrusive program on or off when using the car. The latter variant seems most analogous to a variant of the original Trolley Case in which the driver himself relinquishes the power to make a decision in a dangerous situation and hands this power to a bystander.

Ordinarily, it is only governments or their agents that are permitted to act like intrusive bystanders. And ordinarily when government takes on the role of a fully intrusive bystander it is because it has some duty, not a mere permission, to protect people (e.g., to prevent citizens from harming other citizens). This contrasts with the bystander in the Bystander Case, who is thought to have only a permission rather than a duty to redirect the trolley even when the driver has no power to act.[29] As a private person, perhaps the bystander has no right to deliberately take still-retained power away from a driver without the driver's consent, especially if it is power over his own trolley. (Notice that this is consistent with the bystander permissibly interfering in some other way with the driver's controlling the outcome. For example, suppose the bystander quickly pushes a boulder on the track so that it stops the trolley from hitting the five. When the trolley hits the boulder the trolley is also diverted toward killing another person. This side effect need not make it impermissible for the bystander to act.) On the other hand, since the driver is a threat to five people, why is he not liable to having his power over his vehicle being deliberately removed when this cannot harm him and will ensure that appropriate diversion of the threat takes place?

2.5.3.2. Company Agents

Are there additional reasons why programmers have a right and also should be willing to do what intrusive bystanders may possibly not have a right to do or perhaps should not be willing to do? Arguably programmers have such a right and should be willing to act on it because they are agents of a company that is in a distinctive position of producing cars that may cause harm. On one view this distinctive position implies that if car producers can make cars that in morally permissible ways cause fewer casualties, other things equal, they have a duty to the community to do so. This contrasts with a second view, that the company has duties only to the purchasers of its cars and to its stockholders.[30] It also contrasts with a third view, that the company has none of these duties and is at liberty to make any sort of car it wants. If potential drivers do not like it, they can refuse to buy it, and if the company wants to stay in business they will have to change their product.[31] In what follows, I will consider only what the first view may imply.

Unlike the bystander in the Intrusive Bystander Case, programmers are employed by producers who are (in part) analogous to those who made the trolley that malfunctioned and endangered people in the original Trolley Case. Indeed,

for them the choice may be closer to killing more or killing fewer people rather than letting more die or killing fewer (as it is for the intrusive bystander). Some may say "Cars don't kill people, people do," and indeed in Partial Cases it could be the driver's failure that causes a problem. But even in the latter case, a company might have a duty and also reasonably want their product, other things equal, to cause fewer rather than more deaths at the hands of the driver (e.g., by creating cars that will not start until drivers satisfy an in-car device that tests for alcohol level).

In addition, the company programmer is not intruding after the car has been purchased but ex ante. So he need not be changing what the buyer could expect at the time of purchase. This is also unlike the bystander in the Intrusive Bystander Case who would first get involved at the time that redirection is needed. (This difference would be present even if we imagined that the intrusive bystander was the producer of the defective trolley who had not acted ex ante to reduce deaths.)

If the company is determining what a car that might threaten people through the car's failure alone should be programmed to do, it seems appropriate that they think of themselves as programming *requirements* on how the car should move. This would be comparable to (what many consider) an obligation (not mere permission) of the trolley driver to move the trolley so as to minimize those killed (at least when he would not be harmed). This contrasts with merely providing the car with driver-initiated options for minimizing those killed. Programming requirements seem clearly called for in the Complete Case, where company programmers are deciding what a car that on its own will kill some should be programmed to do so as to kill the fewest in a permissible way.

2.5.3.3. Programmers and Drivers

So far we have considered the role of programmers solely in relation to one possible duty of their company regarding its product: to reduce numbers of people killed. Now consider whether and how this duty may combine with some duties specific to drivers. If the original Trolley Case is a guide, then in the Partial Case the driver would have a duty, other things equal, to redirect so as to minimize deaths, at least when the driver is not at risk of harm. Then programmers who are agents of the company might be in a stronger position to program a car ex ante so as to bring about a death-minimizing outcome because that outcome corresponds to one that would result from a driver's doing his duty. However, there could be conflicts between the driver's duties and the company's duties. For example, suppose that if the driver diverts from killing five, the one he will kill is his child. Presumably he does not have a duty to divert and may even have a duty not to divert. A company program to reduce deaths would divert the car. Even though the driver would not actually be responsible for killing his child if the program diverted the car, it seems wrong for the company to ignore a driver's

moral permission or duty not to divert even when it is the car that malfunctions. Perhaps cars could be programmed to act on such personal information from a driver once a lie-detector device in the car had passed it as reliable. (I will discuss the relation between company's and drivers' permissions and duties further later.)[32]

2.5.3.4. Programmers as Drivers

Another difference between company programmers and the bystander is that the latter is assumed not to be involved as either a driver or as a potential victim. But ex ante those who program cars can also reasonably suppose that ex post, when the car will do as it is programmed, they might be one of those involved in a situation in which people are threatened (e.g., either as the driver or one of those outside the car threatened by it). So unlike the bystander, in the Partial Case they could be helping themselves to bring about an outcome they would have a duty to bring about as a driver. They also stand to benefit (or be harmed) from programming decisions they made because the lives saved (or taken) may be their own. This could affect the prudential rationality for them of programming in a morally permissible way even if it does not result in a moral obligation to program in that way.

2.5.4. Pedestrian Liability

A fourth issue to consider is that in the standard Trolley Cases all the people who might die are thought to be equally innocent, nonthreatening individuals whose actions do not make them deserving of or liable in virtue of their actions to being killed by the trolley. (This is on the continuing assumption that the driver cannot be harmed by any action that helps others.) But suppose five people irresponsibly run in front of a car against a red light. (Call this the Irresponsible Five Case.) This does not make them deserve to be killed, but it might make them liable to being killed rather than one innocent person who would be killed if the car were redirected away from them. Desert and liability are commonly distinguished in the following ways: Giving people what they deserve even if it is something bad is thought to be intrinsically good if it is proportional to what they have done. Doing something bad to someone because his actions have made him liable to have it done is consistent with the bad being out of proportion to what he has done, with it being regrettable that the bad thing must be done to him to prevent harm to someone else not liable to bear it, and preferable that there be another way to achieve this good end.[33]

This Irresponsible Five Case could be one in which a car program should not minimize the number of people killed because not everything else is equal. This is because the degree of liability is a morally relevant difference between the five people and the one other person who would be struck in diversion, making everything not be equal among them.

There might even be a case in which several people are hurled at a stationary car (e.g., by a tornado) and would be killed by impact if the car were not redirected. In this case if they impact the car, harm would come only to them and to no one else. However, it may still be best to treat them as innocent threats because their trajectory, for which they are not responsible, causes a problem that could be avoided only if the driver diverted the car, thus killing a pedestrian. I suggest that if they could divert themselves at some moderate cost to themselves to prevent the pedestrian's dying through diversion, they should do so. Furthermore, it seems that their responsibility to do this is greater than that of a mere bystander to pay the same cost if this would prevent the death of the pedestrian. This is so even though they and this bystander are equally morally innocent of causing the problem situation. People may simply have a duty at some moderate cost to correct the inappropriate location of their body.[34] If they are unable to do so, their having this duty may make it permissible for others to impose at least the same cost on them. It might also be argued that it is permissible for others to allow them to be killed by impacting the car rather than have a driver redirect to a nonthreatening pedestrian who would be killed. This is so if costs they could permissibly be made to bear exceed those that are grounded in their personal duties to make moderate sacrifices.[35]

If these claims are correct, then a company would have a duty in programming to take account of liability to be harmed and moral susceptibility to have harms imposed in addition to any duty to minimize lives lost. (For simplicity, I will here. include both these under an extended notion of liability to be harmed.) Liability might either constrain reducing numbers killed to some degree or possibly have lexical priority over reducing numbers killed. Hence it would be important for a programmed car to be able to detect not only the number of people whose lives are at stake but their degree of liability to be killed. There may be heuristics for detecting this. For example, a car could detect if a pedestrian was crossing against a light or if a driver was speeding. Possibly, probability of liability based on evidence of past differential liability in different circumstances could be used in a program.[36] If cars could not be programmed to detect relative liability, would it be correct to program them at all? Possibly it would be if cases requiring ability to detect liability were rare enough, if drivers were no better than programmed cars at determining relative liability, or if drivers were no more likely than a programmed car to behave on a correct determination of liability.

2.5.5. Driver Liability

2.5.5.1. Principles

As noted earlier, in the standard Trolley Cases neither the driver nor any passengers on the trolley are at risk of being killed. So far in discussing moral issues in Partial and Complete Cases I have been assuming this is so as well. Let us now drop the assumption that the driver cannot be killed instead of some pedestrians and consider his liability to be harmed. As noted in section 2.3, I consider these cases to be Innocent Threat rather than Trolley Problem Cases.

In the course of her 2008 discussion of the Trolley Problem, Thomson claimed that if a nonculpable driver of a car will kill nonthreatening innocent pedestrians unless he redirects, he has a duty to do so even if he is the one who will then be killed.[37] If he has this duty, it is not because he deserves to die or is even at fault. Nor is it simply because sacrificing himself will decrease the number of people killed (if it would), for individuals may have a morally sanctioned personal prerogative not to sacrifice themselves for that goal even if social institutions had to pursue it. Thomson does not explicitly say but she may think the driver's duty (which supersedes his prerogative) arises from his being responsible for setting in motion a car that can kill innocent people.

But note that if the driver has this duty to divert at the cost of his own life because he started the car, then he could have this duty even when only one other person (not a greater number) would be killed by him. Indeed suppose it were possible for two drivers to be responsible for driving the car (or for a passenger giving directions to a driver being jointly responsible with the driver for the car's movements). Then the two of them could have a duty to redirect even if this kills them rather than one innocent pedestrian who would otherwise be hit by their car. So an argument for a driver having a duty to impose the death on himself need not be based on and could conflict with reducing the total number of people killed.

Some may find Thomson's conclusion in her driver case hard to accept because the driver himself has to do what will kill him. A driver who did not do this would most probably be morally and legally excused. However, excuse is not the same as justification, so he may still have failed to do what was right.[38] Furthermore, it would not be as hard for a bystander or a programmer to do what imposes the loss on the driver by either redirecting the car or programming it to redirect, and so the grounds for the driver's excuse would be eliminated. In addition the driver might be liable to have this done to him because he started the car even if he has no duty on these grounds to divert himself and so is even justified in not diverting himself.

Others may reject these conclusions because they think that innocent pedestrians share the liability to be harmed since they are as causally responsible for an accident as the driver simply by being where the car is going; if they were not there, there would have been no accident. I do not accept this view for it seems that a crucial difference between a driver and a pedestrian is that pedestrians are not entities that can damage or harm others on impact in the way cars can harm pedestrians.[39] I acknowledge that this is a complicated issue and moral conclusions could change depending on how this issue is decided.

Some also think that the driver's liability could depend on her reasons for driving. For example, driving an ambulance to fulfill a duty to save other lives might so strongly justify imposing ordinary risk on others that it reduces a driver's asymmetrical liability to suffer any harm relative to pedestrians.[40] But soldiers may have a duty to fight, and yet it is commonly thought they rather than noncombatants (even of the enemy country), who could be considered analogous to pedestrians, should absorb harms in war. It is not my aim in this paper to settle who among the nonrisky, nonnegligent individuals is liable and on what ground. I am primarily concerned to emphasize that wherever it is determined by argument that liability for bearing harm should lie when someone must be harmed, companies could have a duty to take that liability into account in designing programs. This could conflict with other prima facie duties they may have to minimize those killed and to protect the driver.

This raises the practical issue that people may seek to buy cars that are programmed to always favor survival of the driver. (This is so even though when they think of themselves as possible pedestrians on other occasions, they might not favor such a program. In a televised discussion with a philosopher about self-driving cars, a Public Broadcasting Service interviewer, thinking of herself as a driver, said that if there is an accident, she wants to be the one to survive. The philosopher interviewed did not respond that this may sometimes not be the morally correct outcome.) Responding to consumer demand, companies might seek to minimize lives lost and take into account liability to bear harm so long as this applied to everyone besides drivers. Companies could decline to program so as to achieve the morally correct outcome in the light of drivers' duties or liability to bear harm to any greater degree than drivers would ordinarily do so on their own.

If one wanted companies to do more than this, one could try to get them to show that all people would come out better if all drivers were held to programs with higher standards (as in solutions to Prisoners' Dilemmas). Alternatively, one might insist on government regulation to ensure moral solutions that take into account drivers' liability to being harmed and to ensure uniformity in the programs installed. No producers should get a business advantage by providing an "immoral" car that does not incorporate at least "minimal morality" regarding

numbers of lives saved constrained by appropriate liability considerations simply because this will increase their sales.[41] On the other hand, producers and governments should not require the production of what would also be immoral cars, imposing burdens on drivers to which they are not liable and to which they do not consent, for the sake of always maximizing lives saved despite the liability of pedestrians. However, producers should not necessarily be prohibited from programming at the request of drivers altruistic (or supererogatory) cars that allow drivers to bear burdens beyond both those they owe (which a "strictly dutiful car" could ensure) or which may permissibly be imposed on them.

Finally, note that it is easier for a driver to *buy* a car that she knows is programmed to do what will sacrifice her than to actually sacrifice herself; buying such a car is not a way to sacrifice oneself. This is in part because it is uncertain whether one will ever be in a situation where one would be sacrificed by one's programmed car. Combined with an ex ante desire to do what might be one's duty in a tragic case and foreseeing one's lack of courage to do it at high cost, some drivers may actually want to buy a car that is programmed to produce outcomes that track either their duty or liability to bear harms (where these differ). (This is in addition to such drivers considering the possibility that they will at times be pedestrians.)

2.5.5.2. Illustrative Cases

To reinforce these conclusions consider some cases.

2.5.5.2.1. Case 1

Suppose (for the sake of argument) an in-control driver would have a duty to sacrifice himself by diverting rather than kill innocent nonthreatening pedestrians. Does this imply that company programmers have permission or even an obligation to program a threatening car in the Partial Case to divert from killing more nonthreatening pedestrians to killing fewer driver(s) of the threatening car? It would not be surprising if the driver couldn't be trusted when he is in control of the car to do his (assumed) duty to minimize lives lost when doing so would cost him his life. So without forcing anyone to sacrifice himself, they would arrange for the car to generate an outcome that corresponds to that from the performance of a duty that (arguably) *both* the producer and the driver have in this case to minimize lives lost even at the driver's expense.

Suppose the driver had no duty to sacrifice himself. He still might be liable to have costs imposed on him by others. Suppose he was no more liable in virtue of driving to bear costs than anyone else. In diverting the car when this kills him the programmers would still treat him no worse than they (or he) would treat an innocent pedestrian in diverting to her to minimize innocent pedestrians killed,

though she has no duty to sacrifice herself and is not liable to bear harm in virtue of her actions.

2.5.5.2.2. Case 2

Suppose a driver would have a duty to divert at the cost of his own life to save pedestrians. Does this imply that the company should program to divert a car from killing fewer (or the same number of) nonthreatening pedestrians in a way that results in the death of more (or the same number of) driver(s) of the threatening car? If programmers did this, they would not act on any producer's duty to minimize lives lost due to its machine. Their disfavoring the driver(s) would have to involve either arranging for an outcome because it corresponds to one that would result if a driver performed his duty to sacrifice himself and/or involves imposing harm to which the driver is liable even in the absence of his duty. Acting on this consideration would override the producer's other prima facie duty to minimize lives lost since more lives might then be lost. If a company has a duty to take account of liability to bear costs in programming cars, then while it may be contentious that the driver is liable if he is, the company has its own duty to take account of his liability even if this does not minimize lives lost.

2.5.5.2.3. Case 3

Should the company program to *prevent* the diversion of a threatening car from killing both one innocent nonthreatening pedestrian and the driver of the car toward killing two innocent nonthreatening pedestrians? In this case the driver would lose his life not in being diverted but in not being diverted; his life would be saved by diverting toward more pedestrians. The same number of people would be killed either way.

This case raises the following question: When costs to a driver would be high, could there be a moral difference between (1) a program preventing a driver from increasing the number of pedestrians killed and (2) a program reducing the number of pedestrians a driver kills? Suppose there is such a moral difference, in favor of (1). Then it may be permissible for a program to at least prevent the driver's life being saved by diverting when more pedestrians will be killed (though the same number of people would be killed). This could be so even if the program should not lethally divert the driver in order to prevent his killing these pedestrians. Hence if he were headed to killing two people but would survive this, such a program would not divert him toward one other pedestrian when he would die. However, the program would prevent his diverting from killing one pedestrian to killing two, though he will not survive without the diversion. (If the driver's life had the same moral weight as a pedestrian's, preventing his diversion in the latter case would yield the same outcome as not turning from killing one set of two people toward killing another set of two people.)

However, suppose the driver would kill two pedestrians and himself if the car is not diverted, and he would be saved but kill two other people if the car is diverted. In this case, if the car is diverted, an additional life (of the driver) would be saved and there would be no increase in the number of pedestrians killed. Even if there is something to be said against killing two other people rather than letting those originally threatened be killed, saving the life of a driver who does not deserve to die seems important enough to justify diversion. (If the driver himself were in control, he could permissibly save his life in this way since he wouldn't be increasing the number of pedestrians killed.)

2.5.5.2.4. Case 4

Should the company program to divert a car from killing more nonthreatening pedestrians in a way that kills the single driver of the car rather than in an alternative way that kills one different nonthreatening pedestrian? In this case, the car producer's duty to minimize numbers killed would be satisfied either way. If the driver is liable to bear harm, the producer should arrange for the car to divert in a way that kills the driver rather than in a way that kills a different pedestrian.

The overriding conclusion of the discussion of cases 1–4 is that even if producers should not bring about outcomes simply because they correspond to the performance of drivers' duties, bringing about the same outcome or even one more burdensome for the driver can sometimes be necessary in order for producers to carry out their own duty to take account of a driver's liability to bear harm.

2.6. Passengers and Other Drivers

In the two standard Trolley Cases (with a driver or a bystander called on to redirect), the five potential victims and the one to whom the trolley could be redirected are not imagined to be either on the threatening trolley or on another trolley. But programmers for cars recognize that a car can face a collision with other cars and that people initially and potentially threatened might also be in the cars. Rampolla and MIT's Moral Machine website present cases in which if redirection occurs, nondrivers who would be harmed are passengers in the threatening vehicle. I do not think they deal with cases in which those initially threatened are also in vehicles or those potentially threatened are in vehicles other than the initially threatening vehicle. Let us consider a variety of such cases.

2.6.1. Other Drivers

Suppose for the sake of argument that a driver is liable to be sacrificed relative to nonthreatening pedestrians she would otherwise hit simply because she is

driving a threatening car. Are drivers in nonthreatening cars also liable to have harm due to a threatening car redirected to them rather than to nonthreatening pedestrians? For example, suppose the driver in the car threatening pedestrians cannot prevent harm to them by herself bearing the cost. She or the program running her car can either redirect toward another car with a driver or to another pedestrian. Choosing the car with the driver might be justified by analogizing the case to players in a dangerous game (in this case, car driving) who should when possible confine themselves to injuring one another rather than nonthreatening nonplayers if someone must be injured. If so, cars should be programmed to detect and at least sometimes redirect to other cars rather than to pedestrians even when those cars have drivers and diverting in this way will not minimize deaths. "Playing the game of driving" would then be another source of liability to harm.

When a threatening driver could bear costs to prevent harm to others, additional issues related to liability to bear costs will arise: Should the number of all drivers' lives at stake in a decision, possible fault, or merely who is the initial threatening driver determine programming? Could the programmed car detect and "act" on those factors at least as well as unassisted drivers?

Perhaps in some cases involving multiple vehicles it may be possible to distinguish between something like an offensive and a defensive threat. An "offensive" innocent threat would be presented by the Partial or Complete car that initially nonintentionally and nonnegligently threatens either pedestrians or other drivers. A defensive threat might be presented by a vehicle that has to respond to that initial threat; doing so may result in its threatening either the initial threat, other vehicles, or nonthreatening pedestrians. Even if an offensive threat should be programmed so that its driver is sacrificed rather than another driver to whom he presents a threat, the driver who becomes a defensive threat may be liable to be harmed rather than pedestrians his vehicle might harm. (This is because of his participation in the dangerous practice of driving.)

2.6.2. Passengers

Aside from the drivers of threatening vehicles, and even in Complete Cases where drivers are absent, there may be passengers in vehicles.[42] Suppose nondrivers in the threatening vehicle would be killed if the car were redirected from harming nonthreatening pedestrians. Should vehicles that can detect the presence of people inside the car be programmed to count their lives on a par with pedestrians threatened by the vehicle and do what reduces the number of people killed?

People often voluntarily decide to be passengers in a vehicle that they know potentially threatens pedestrians. Furthermore, the vehicle might not have

started at all in the absence of passengers, and passengers may tell vehicles where they want to go if not how to get there. For example, one of the benefits of completely self-driving cars is that they would increase the mobility of blind and paralyzed people. But if such passengers gave a command for the car to start, are they not like the driver who started the trolley that then went beyond his control (even if the passengers never drove the car to begin with)? Suppose such drivers should be given less weight than pedestrians because they are liable to bear costs. Then shouldn't passengers whose directions start a car also be given less weight than pedestrians by a program for Complete Cars?

What about voluntary passengers who did not start the car but chose to join those who did start the car? Their joining a dangerous game provides a ground for some liability to bear costs.[43] When a car with seven passengers threatens two pedestrians perhaps, each of the seven should count for only a fraction of a person in a calculation of lives lost. However, joining oneself to a car before it becomes a threat is still different from, for example, hopping a ride on what one knows is a vehicle headed to killing pedestrians. In the latter case, the passengers' lives should certainly have reduced weight relative to the lives of pedestrians when deciding whether to divert the vehicle for they knowingly attach themselves to a threat that should have been diverted. If possible, a car should be programmed to detect such morally relevant differences among passengers.

2.6.3. Mere Cars

Do the following conclusions at least seem certain? When a car with a driver and/or passengers threatens a completely empty car, the empty car should be destroyed rather than kill pedestrians or other drivers and passengers. Also, empty Complete Cars should be programmed to "sacrifice themselves" rather than pedestrians or drivers and passengers in other cars even when the latter are causally or (sometimes) even morally responsible for the initial problem. These conclusions seem to follow from the view that lives of persons take precedence over property, at least when there is no intentional wrongdoing by those people that aims at destruction of property.

However, what if the empty Complete Car is the only one that can be sent to save many other people? I think it (like the Ambulance) cannot be allowed to run over even fewer nonthreatening pedestrians in order to avoid its own destruction and continue on its mission. But are those who are morally responsible for being in harm's way (like those in the Irresponsible Five Case) liable to being harmed by the car rather than having it destroyed when it is necessary to save many other lives? Being liable to bear costs rather than have the car redirected to *kill* others,

as is true of the five in the Irresponsible Five Case, is still morally different from being liable to bear costs so that the car can go on to *save* others, I think.[44]

Notes

1. F.M. Kamm, *Morality, Mortality*, vol. 2 (New York: Oxford University Press, 1996).
2. I am assuming that the Partial Cases involve taking control away from a driver who is not allowed to drive only in certain dangerous situations. Other types of Partial Cases may involve self-driving cars that turn over control to the driver only in certain dangerous situations. In addition, it is said that self-learning cars need not be programmed. For simplicity's sake, I will speak of cars that decide what to do as acting on a program.
3. Though the case was created by Philippa Foot in "The Problem of Abortion and the Doctrine of Double Effect," in *Virtues and Vices* (Los Angeles: UCLA Press, 1978): 19-33, it was only later called the Trolley Problem in Judith Thomson, "Killing, Letting Die, and the Trolley Problem," *Monist* 59, no. 2 (1976): 204–17.
4. See Judith Thomson, "The Trolley Problem," *Yale Law Journal* 94 (1985): 1395-1415. In a switch from her 1976 article, Thomson came to apply the term "Trolley Problem" to this case alone, though others did not.
5. In another type of case, which I call Crosspoint, a bystander must decide whether to turn a trolley that will kill many if it remains at a crossing point toward killing five or killing one. Here the bystander is choosing between letting many die or killing either five or one.
6. Such a case was also introduced by Thomson, who called it the Fat Man Case in her "Killing, Letting Die, and the Trolley Problem."
7. See Chris Rampolla, "The Trolley Problem Reimagined: Self-Driving Cars," *Aero*, March 31, 2017. I was directedto this site by a nonphilosopher professor who works on the Trolley Problem and who saw it as popular discussion that provided a good introduction to the problem for the general public.
8. Rampolla actually speaks of the car, not the driver, striking the pedestrian and the driver possibly causing a collision that saves the pedestrian but will likely kill him. Reserving the term "killing" only for the latter effect seems a biased description.
9. The problem of applying results achieved by assuming certainty when real life presents us only with risks is a topic discussed by others. See note 25.
10. Neither a nonphilosopher professor nor a postdoctoral psychologist working on the Trolley Problem, both of whom were present when I discussed Rampolla on April 23, 2018, had ever heard of Innocent Threat Cases. Perhaps they would still have realized that different moral problems might be raised by them from those raised by the standard trolley cases. However, Joshua Greene, a philosopher and psychologist who has written about the Trolley Problem, gave as a real-life example of it (when speaking at a Safra Ethics Center conference dinner at Harvard in 2017) a case in which doctors must decide whether to confine a person carrying a contagious disease, thus

imposing costs on him in order to save others from the disease. He reported that doctors were concerned about imposing costs on one person to save others, and he gave the impression that this was like the concern about turning the trolley on one person to save others. But in the medical case costs would be imposed on the person who presents the threat and doctors should not, I think, be as concerned about imposing a cost on an innocent threat as on an innocent nonthreatening person.

11. See Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974).
12. See Saba Bazargan, "Killing Minimally Responsible Threats," *Ethics* 125, no. 1 (2014).
13. Though killing the driver might stop the threat, as when killing him causes his body to fall on a brake that is otherwise inaccessible to him.
14. Prof. Bert Huang refers to a psychological study on trolley problems in which researchers expose subjects to "the actual argument—that no one is obliged to sacrifice his own life to save others, and that it seems immoral to force another to make a sacrifice one would not have to make oneself." B. Huang, "Law and Moral Dilemmas," *Harvard Law Review* 130 (2016): 673. This description suggests that the person who would die from the trolley if it is redirected toward him in the standard cases is made to sacrifice his life. But this ignores the possibility that there is a conceptual (and moral) difference between imposing loss of one's life on oneself (either voluntarily or coerced) and having that loss imposed on one by another.
15. In sum, the following are among the cases we have so far distinguished: (1) someone who threatens but cannot be threatened decides on which nonthreateners to impose losses (the standard driver Trolley Case); (2) a bystander who cannot be threatened decides whether to impose losses on nonthreateners with no possibility of harming the threatener (the Bystander Case); (3) someone who threatens and can be threatened decides whether to impose losses on himself or his potential nonthreatening victims (an Innocent Threat Case as I use the term); (4) a bystander who cannot be threatened decides whether to impose losses on a threatener or allow (or cause) losses to be imposed on nonthreatening victims (another Innocent Threat Case).
16. This is also true in many cases presented by the MIT Moral Machine website, http://moralmachine.mit.edu.
17. For objections to Foot's proposal, see F.M. Kamm, *The Trolley Problem Mysteries* (New York: Oxford University Press, 2015).
18. Marya Zilberberg, "Medicine as the Trolley Problem," *Healthcare, etc.*, July 25, 2012, http://evimedgroup.blogspot.com/2012/07/medicine-as-trolley-problem.html.
19. Bioethicist Nir Eyal once suggested this (much before mounting evidence about the dangers of e-cigarettes). The example used merely as a hypothetical case remains useful for my purposes.
20. The Ambulance Case with which I began this discussion also contrasts with diverting the trolley. Though it involves choosing between saving five (by getting them to the hospital) and killing one, it is not like the Bystander Case (which also involves choosing between saving five and killing one). It may seem that one way of marking the distinction is that, unlike the Trolley, the Ambulance (like the Gas or Bomb) that would kill the one is a new threat that is not already threatening the five. But the case (presented in the text) in which the diverted trolley causes a new rockslide that kills

one person shows why it is not quite right to say that the problem in the Ambulance Case is that a new threat kills the one. It would be better to say the following: It is not to remove a threat that the ambulance itself presents to the five that we would consider having the ambulance continue on killing one other person, and so it is impermissible to drive the ambulance over the person on the road. However, elsewhere I have considered what I call the Lazy Susan Case to be like the Trolley Case, though it does not even involve redirecting a threat from the five. Rather it involves moving people away from a threatening trolley that cannot be redirected. In one version, their being moved will create a new rockslide that kills a bystander. I have argued that doing what saves the five but kills the one is as permissible in this case as in the standard trolley cases, even though the one person is killed neither by what threatens the five nor by a new threat created by the threat itself moving away from the five. I argued that the Lazy Susan Case is like the Trolley Case because the same principle explains the permissibility of both turning the trolley and turning the Lazy Susan. In my view, this principle (put roughly and far too simply) is that what kills one other person just is the five being saved. The trolley turning away which kills the one does not merely cause the five to be saved (as the bomb in the Bomb Case would); it constitutes the five being saved. Similarly the five being moved away from the trolley on the Lazy Susan, which leads to one other person dying, just is the five being saved. For further discussion of this, see F.M. Kamm, *Intricate Ethics* (New York: Oxford University Press, 2007) and *The Trolley Problem Mysteries* . I omit further discussion of Lazy Susan–type cases here since they do not seem pertinent to real-life cases of self-driving cars.

21. I do not in this chapter defend the correctness of these judgments and the principles that justify them (though I said something about this in note 21). Thomson herself came to reject her earlier judgment that it is permissible for the bystander to redirect the trolley, thus killing one and saving five. See Judith Thomson, "Turning the Trolley," *Philosophy & Public Affairs* 36 (2008), 359-374 and my discussion of her later view in *The Trolley Problem Mysteries*.
22. Principles and rules are commonly distinguished. For example, H. L. A. Hart conceived of the legal system as consisting of rules. By contrast, Ronald Dworkin thought that system fundamentally consisted of principles (such as "not benefiting from one's crime" or "fair play") that required more interpretation than rules and that could ground rules. The principles might also guide us when rules run out or when rules lead to conclusions in particular cases that are inconsistent with grounding principles.
23. Fiona Woollard and Will McNeill discuss these issues in their "Driverless Cars" and "Ethics without Algorithms," both unpublished manuscripts.
24. I continue to put to one side the important issues of certainty versus probability of deaths and knowledge of this. In the standard hypothetical trolley cases one assumes certain death for the five if the trolley continues or for someone else if the five are saved. One also assumes knowledge of this by the decision-maker. This is not necessarily true in real life. On this problem of applying Trolley Problem reasoning to programming cars, see Sven Nyholm and Jilles Smids, "The Ethics of

Accident-Algorithms for Self Driving Cars: An Applied Trolley Problem?," *Ethical Theory and Practice* 19, no. 5 (November 2016): 1275–89.

25. Note also that in trolley cases it is because the trolley is no longer under the driver's control in heading to the five (though it is also not self-driving) that a problem initially arises. By contrast, it is because human drivers in cars completely under their control often fail to do the right act that initial life-threatening situations often arise. It is somewhat ironic that (1) cases in which a problem arises in the first instance because a vehicle lacks a driver in control are being looked to (by some) for guidance about what to do (2) when a car's lacking a driver in control is supposed to prevent problems from arising in the first instance.
26. See, for example, Thomas Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986).
27. For defense of a view like this see, for example, my *Morality, Mortality,* vol. 2
28. I am grateful to Jesse Berthold and Arthur Applbaum for raising questions that led to some of my responses in this section.
29. This may be because the bystander will wrong the one person he kills even if he acts permissibly in doing so.
30. Suggested by Larry Temkin.
31. Suggested by Shelly Kagan.
32. We could also consider the role of programmers' duties in relation to possible victims' duties. For example, suppose the five toward whom the car is headed were the parents and guardians of the one person toward whom the car would be diverted. They might have a duty or preference to see to it that the car is not diverted. However, drivers are no more likely to know of such relations between potential victims than programmed cars would, whereas they do know about their own duties and preferences.
33. On the distinction between desert and liability, see Jeff McMahan, *Killing in War* (Oxford: Oxford University Press, 2009). In the Irresponsible Five Case, McMahan would say that the five are liable to the harm because they have "assumed the risk." A crucial issue is whether being liable should be a tiebreaker when all else is equal between potential bearers of a loss or whether it should come with built-in limits on the loss to which one can be liable depending on what makes one more liable than someone else. On the latter view one might be liable to a higher chance of bearing a loss (e.g., 80%) or liable to bearing a loss only up to size x. If someone must bear a certain-to-occur loss or a loss over x, then on this second view who should bear the loss should be determined by giving equal chances. On the first view, being liable wouldn't have such built-in limits and could serve as a tiebreaker that determines on whom the certain-to-occur loss or loss larger than x should be placed. (The number of people harmed might count in determining the size of the loss.)
34. I discussed cases of this sort in F.M. Kamm, "The Insanity Defense, Innocent Threats, and Limited Alternatives," *Criminal Justice Ethics* 6, no. 1 (1987): 61–76.
35. The permissibility of turning the trolley on one innocent person relies on the view that what may be imposed on someone exceeds his duty to impose harm on himself. But if liability depends on a person's action or movement, there would not be this ground for imposing harm on the one person in the trolley case. Furthermore, more

people who are liable to have harm imposed on them might sometimes permissibly be harmed to save fewer people. By contrast turning the trolley and harming the one nonliable person depends on fewer people overall dying.

36. I owe these suggestions for the heuristics to Jeff McMahan, Shelly Kagan, and Larry Temkin.
37. Thomson discusses this case in her "Turning the Trolley," 369.
38. Note that his duty is not necessarily to actively redirect when that would have killed him. For there might be a case in which unless he diverts he will continue on in a way that results in his being killed, but if he diverts to save himself he will kill others. Then he might have a duty to refrain from diverting. (I discuss such a case later in the text.)
39. However, there is the difficulty of distinguishing morally between a moving pedestrian who is walking and an innocent hurled at a car. Why would the latter be liable to bear costs, as I argued earlier, and not the pedestrian since neither directly threatens harm or damage to others?
40. Jeff McMahan holds such a view.
41. Analogously, suppose a college complained that it couldn't attract students if it did not allow some cheating because other schools allowed some cheating and students preferred those schools. The solution is not to give up the correct moral standard but for all schools to agree to enforce the standard. Suppose students would then prefer no education (comparable to people not buying any cars that tracked moral requirements). Suppose this education outcome was bad (as might not be true if as a consequence of not buying cars people used only public transportation). Then it might be necessary to either require the practice of education or make it more attractive in some way other than by allowing some cheating.
42. Unlike the Moral Machine website, I shall consider only person passengers, not nonperson animal passengers. That website also considers animal pedestrians.
43. It might be said that passengers (and drivers) stand to benefit from using cars and this is what grounds their liability to be harmed rather than pedestrians. In the case of both passengers and drivers I do not wish to derive liability to harm from standing to benefit. For even if drivers and passengers did not stand to benefit from using cars, the risk of harm to others from the devices they use is an important reason why they might be liable to bear costs when someone must.
44. I am grateful to Mathew Liao for inviting me to write this chapter. I am grateful to him, Shelly Kagan, Jeff McMahan, Larry Temkin, students in my Rutgers philosophy seminars, and audiences at the Edmond J. Safra Ethics Center of Harvard University and at the University of Granada Philosophy Department for comments on earlier versions of this chapter.

References

Bazargan, Saba. "Killing Minimally Responsible Threats." *Ethics* 125, no. 1 (2014): 114-136.

Foot, Philippa. "The Problem of Abortion and the Doctrine of Double Effect." In *Virtues and Vices*. Los Angeles: UCLA Press, 1978: 19-33.

Huang, B. "Law and Moral Dilemmas." *Harvard Law Review* 130 (2016): 659-699.
Kamm, F.M. "The Insanity Defense, Innocent Threats, and Limited Alternatives." *Criminal Justice Ethics* 6, no. 1 (1987): 61–76.
Kamm, F.M. *Intricate Ethics.* New York: Oxford University Press, 2007.
Kamm, F.M. *Morality, Mortality.* Vol. 2. New York: Oxford University Press, 1996.
Kamm, F.M. *The Trolley Problem Mysteries.* New York: Oxford University Press, 2015.
McMahan, Jeff. *Killing in War*. Oxford: Oxford University Press, 2009.
Nagel, Thomas. *The View from Nowhere*. New York: Oxford University Press, 1986.
Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.
Nyholm, Sven, and Jilles Smids. "The Ethics of Accident-Algorithms for Self Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Practice* 19, no. 5 (November 2016): 1275–89.
Rampolla, Chris. "The Trolley Problem Reimagined: Self-Driving Cars." *Aero*, March 31, 2017.
Thomson, Judith. "Killing, Letting Die, and the Trolley Problem." *Monist* 59, no. 2 (1976): 204–17.
Thomson, Judith. "The Trolley Problem." *Yale Law Journal* 94 (1985): 1395-1415.
Thomson, Judith. "Turning the Trolley." *Philosophy & Public Affairs* 36 (2008): 359-374.
Woollard, Fiona, and Will McNeill, "Driverless Cars" and "Ethics without Algorithms" (unpublished).
Zilberberg, Marya. "Medicine as the Trolley Problem." *Healthcare, etc.*, July 25, 2012. http://evimedgroup.blogspot.com/2012/07/medicine-as-trolley-problem.html.

3

The Moral Psychology of AI and the Ethical Opt-Out Problem

Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan

3.1 Introduction

Most people are happy to use technology driven by artificial intelligence (AI), as long as they are not fully aware they are doing so. They enjoy music recommendations, email filters, and GPS advice without thinking too much about the machine learning algorithms that power these products. But as people let AI-driven technology take an ever greater place in their lives, they express anxiety and mistrust about things labeled AI. Leaving aside fears of superintelligent robots lording over humanity,[1] only 8% people would trust the mortgage advice offered by an AI program—a shade lower than the 9% who would trust their horoscope for investment advice.[2]

Of course, shopping recommendations and GPS routes arguably do not have a critical impact on people's life outcomes. AI-driven technology, though, is progressively extending into realms in which it will have such an impact, and thus make decisions that fall in the moral domain. Self-driving cars will need to make decisions on how to distribute risk among road users; organ donation algorithms prioritize who will get a transplant; and algorithms already advise judges about who should get probation, parole, or a longer jail sentence.

All these decisions inescapably incorporate ethical principles and complex moral trade-offs. Should self-driving cars always strive to minimize casualties, even if it means sometimes sacrificing their own passengers for the greater good? Should children always have priority for organ transplants, even when an older patient is a better genetic match for an available organ? Should sentencing algorithms always seek to minimize rearrest, even if this minimization results in an unfair rate of false alarms for black and white defendants?

It is not always clear who should be consulted to answer these questions. Should we seek responses from ethicists or listen to laypersons' opinions? Even though ethicists do not necessarily behave better than laypersons, and even though their initial intuitions may not be better than that of laypersons, their training allows them to think more deeply about these questions and provide

Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, *The Moral Psychology of AI and the Ethical Opt-Out Problem*
In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0004

solid justifications for their conclusions. Laypersons' intuitions, in contrast, are often untrained and uninformed.

It would be tempting, then, to discard laypersons' intuitions and preferences about the complex ethical issues raised by algorithms and AI-driven technology. But that would be a grave mistake. To understand why, one must realize that if people are not satisfied with the ethical principles that guide moral algorithms, they will simply *opt out* of using these algorithms, thus nullifying all their expected benefits.

Self-driving cars provide the starkest example of the effect of such an opting-out. Imagine (for the sake of the argument) that some ethicists would agree that self-driving cars should always strive to minimize casualties under a veil of ignorance—that is, that self-driving cars should always take the action that minimizes harm, even if this action is dangerous for their own passengers. This would seemingly guarantee the greatest safety benefits for all road users, measured by the smallest overall number of traffic fatalities. But it would also mean that self-driving cars might autonomously decide to sacrifice (or at least imperil) their own passengers to save other road users—and this possibility is so aversive to consumers that they might opt out of buying self-driving cars, thus forfeiting all their expected safety benefits.[3]

In other words, even if ethicists were to agree on what they believe to be the best ethical principles to guide a moral algorithm, their work would be made null and void if many laypersons were to strongly disagree with them, to the point of opting out of using the algorithm. This ethical opt-out can take several forms. People opt out of using self-driving cars by not buying them. People opt out of organ donation by either not registering as donors or registering as nondonors. Finally, people can opt out of judicial algorithms by electing state court judges who vow not to use such algorithms (in the United States), or by turning to alternative, community-based justice such as sharia councils (in the United Kingdom).

One may still argue that if ethicists were in fact able to come to a consensus about the normative principles guiding moral AI in a given domain, then laypersons should be educated rather than listened to. In other words, the best way forward would be to persuade laypersons by rational argument[4] or implicit nudging[5] rather than to adjust the principles to make them closer to what laypersons spontaneously find acceptable. As a matter of fact, we are agnostic when it comes to this debate. What we note is that *whichever way is actually taken,* public policy will require understanding what people find acceptable—whether with the aim of coming closer to their preferences or of persuading them that their preferences should be abandoned.

In sum, many benefits of AI technology require people to opt into an *algorithmic social contract*, an agreement between citizens, mediated by

machines.[6] To facilitate such agreement, we must understand what principles people expect moral AI to follow, lest they opt out from using, enabling, or allowing beneficial AI-driven technology. And we need this understanding regardless of whether we think people should be educated or accommodated. The problem, then, is how we can achieve this understanding. Here we can draw inspiration from the tools and techniques developed in the field of moral psychology. However, applying these tools to the field of moral AI raises methodological as well as second-order ethical challenges, which we now address in turn.

3.2 Methodological Challenges

Assessing moral preferences is a complicated matter, one that has drawn in not just the field of moral psychology[7] but also subfields of experimental economics and human behavioral ecology.[8] Moral preferences are fluid, multifaceted, and nuanced. To measure them is to accept that much complexity is lost in the measurement, and that some measurement techniques inevitably amount to presenting people with highly simplified, stylized problems—problems that sacrifice realism in order to cut at the joints of moral preferences. Different domains of application call for different degrees of such simplification, as we consider in this section using three examples: autonomous vehicles, kidney paired donation, and algorithmic sentencing.

3.2.1. Autonomous Vehicles

The most famous stylized moral dilemma is known as the Trolley Problem.[9] In its most common version, the Trolley Problem presents people with a scenario in which a trolley car is barreling down on five persons, with no time to stop. If nothing is done, these five persons will die. The only way to save these persons is to pull a lever that will redirect the car onto another line. One person, though, is currently on that line and will be killed by the car, with no time to react. The question is whether it is morally acceptable (or even obligatory) to pull the lever.

This specific scenario is frequently criticized as unrealistic. How many times did such a situation actually occur in the real world? Why can't the car just stop? Why are these people standing there instead of walking a few steps, away from harm? These are all legitimate questions, but experimental psychologists (and experimental philosophers, for that matter) simply ask people to accept the premises of the problem in order to discover fundamental principles and processes underlying moral judgment. As a result, the Trolley Problem has led to

many important insights about human morality, despite (or thanks to) its unrealistic simplicity.

Consider now the AI version of the Trolley Problem, in which an autonomous car is barreling down on five persons and cannot stop in time to save them. The only way to save them is to swerve into another pedestrian, but that pedestrian would then die. Is it morally acceptable (or even obligatory) for the car to swerve? This scenario is clearly as unrealistic as the classic trolley scenario. Why is the car driving at unsafe speed in view of a pedestrian crossing? And why are the only options to stay or swerve—surely the sophisticated AI that powers the car should be able to come up with other solutions?

Just like the Trolley Problem and most experimental stimuli in the behavioral sciences, this autonomous car dilemma is a *model*, not a *reflection* of reality. To borrow a turn of phrase, it is meant to be taken seriously without being taken literally: it captures the gist of many genuine ethical trade-offs that go into the algorithms of autonomous cars, and does so in a way that laypersons can understand.

In the real world, every complex driving maneuver influences relative probabilities of harm to passengers, other drivers, and pedestrians.[10] A car that is programmed to favor a certain set of maneuvers may thus have a higher probability of harming pedestrians and a lower probability of harming passengers. Though these maneuvers may only minutely shift the risk profile for any individual, the trade-offs that are being made will become apparent when aggregating statistics over thousands of cars driving millions of miles. And these statistics will prompt the same questions as the stylized dilemma does.[11] For example, imagine that accidents involving one car have a 1-to-2 ratio of passenger to pedestrian fatalities, while another car exhibits a 1-to-7 ratio. Will society accept this discrepancy? Will consumers flock to the second car? Should regulators intervene? Note that we have been there before. For example, "killer grilles" (also known as "bull bars") were banned by many regulators because they disproportionately harmed pedestrians and passengers in other vehicles. Regulators identified the ethical trade-off embedded in a physical feature of the car and acted in the interest of all stakeholders. Should they do the same for the ethical trade-offs embedded in self-driving car software?

By capturing ethical trade-offs embedded in software in a form that all people understand immediately, the stylized dilemma empowers them not to leave ethical choices in the hands of engineers, however well-intentioned these engineers are. To dismiss the stylized dilemma as an abstract philosophical exercise is to hide ethical considerations where lay individuals cannot see them. Most would agree that ethical algorithms should be developed transparently—but transparency is useless if the trade-offs are too obscure for the public to understand.

Stylized dilemmas like the Trolley Problem have a critical role to play to overcome this psychological opacity.

The need for stylized dilemmas should accordingly be assessed as a function of the complexity of the domains to which we apply moral AI. In some domains, it might be possible to measure moral preferences using problems that are actually very close to the real thing. In the next section, we consider one such domain, organ transplants.

3.2.2 Kidney Paired Donation

Too frequently, candidates for kidney donation have access to a living donor who unfortunately is a poor match for them. To optimize the efficiency of kidney allocation, kidney paired donation (KPD) consists of entering candidates and donors in a database, which is then fed to an algorithm that seeks two-way, three-way, or complex chains of donations such that as many candidates as possible find a compatible donor.

The algorithm does not only seek to maximize the number of donations, though. It also uses a scoring rule to determine the priority of each donation (see later discussion) in order to find chains that maximize the number of high-priority donations. While the chain-seeking part of the algorithm might be too complex for laypersons to understand, the same is not true of the scoring rules that determine the priority of each donation. Most criteria in these scoring rules can be readily understood, and the trade-offs they imply may be explained almost straightforwardly to citizens, and to potential donors in particular.

Consider, for example, the criteria shown in Table 3.1, together with the priority points they get under two scoring rules. While the interpretation of the zero-antigen mismatch criterion and the controversies surrounding its use are perhaps best left to specialists,[12] the other criteria are easy enough for laypersons to understand. Three of the criteria are straightforward (travel distance, recipient's age, recipient's prior donor status). The Panel Reactive Antibodies (PRA) score indicates the proportion of the population against which the candidate is immunized, which accordingly restrict the pool of potential donors for this candidate. A candidate with a PRA score of 80 is thus unable to receive a kidney from 80% of donors.

With this information laypersons can readily assess some of the trade-offs implied by the scoring rules, as well as some of their problematic aspects. Consider the problems raised by using cutoffs for continuous criteria such as age and PRA. Why would a five-year-old candidate receive more points than a six-year-old candidate, while the six-year-old candidate does not receive more points than a

Table 3.1 Examples of criteria used in the kidney allocation algorithms of the Alliance for Paired Kidney Donation (APKD) and the Kidney Paired Donation program of the Organ Procurement and Transplantation Network (OPTN), before their 2016 update. PRA = Panel reactive antibodies.

		APKD	OPTN
Zero-antigen mismatch	Yes	6	200
High PRA	PRA > 80%	10	125
	PRA > 50%	6	0
Travel distance	Same region	0	25
	Same center	3	25
Pediatric recipient	Age < 6	4	100
	Age < 18	2	100
Prior donor	Yes	6	150

seven-year-old candidate? Is it fair that a candidate with a PRA of 80 gets a massive point gain compared to a candidate with a PRA of 75, while a candidate with a PRA of 98 receives the same number of points as a candidate with a PRA of 80? These are questions that laypersons can easily understand without the need for researchers to invent stylized dilemmas.

Consider now the relative importance of the criteria and the fact that they can largely differ between the two scoring rules. Why is it that under the rules established by the Alliance for Paired Kidney Donation, being in the same center as the donor awards slightly more points than being seventeen years old, while being seventeen awards four times as many points as being in the same center under the rules set by the Organ Procurement and Transplantation Network? The fact that the scoring rules can largely differ is a telltale sign that we are dealing with fluid, controversial moral trade-offs. And, again, the palatability of these trade-offs is likely to influence people's decisions to participate as donors. Moral psychology can assess the public perception of these trade-offs through experimentation,[13] without the need for simplifying the problem to the extent it had to simplify autonomous vehicle (AV) ethics into trolley problems.

3.2.3. Algorithmic Sentencing

There are other application domains, though, in which the ethical trade-offs are not only hard to explain but also hard to stylize—and these domains will likely

prove the most difficult to investigate with the methods of moral psychology. This is especially the case with algorithmic sentencing. Many US courts now offer judges the option of using an algorithm that provides a risk score for the defendant—for example, the risk that the defendant will not show up at trial (which can lead a judge to decide that the defendant should await trial in jail), or the risk of recidivism or violent crime (which can lead to a longer jail sentence or a sentence in a higher security prison). While there are dozens of such algorithms, some of them created by nonprofit organizations, the best-known exemplars are proprietary algorithms created by for-profit organizations, such as the COMPAS tool created by Northpointe (now Equivant). The opacity of these proprietary algorithms obviously imposes limits on the realism of any experimental vignette; if we do not even know which parameters the algorithm uses, we cannot investigate the public perception of the trade-offs between these parameters.

There are some ethical trade-offs we can experimentally investigate, though, even without knowing the specific implementation of the risk assessment algorithms—but these trade-offs hardly lend themselves to a one-sentence explanation, or to a trolley-like stylized dilemma. To illustrate, let us unpack the controversy that arose about the potential racial biases of the COMPAS tool.

In May 2016 the investigative news organization ProPublica published a story by Julia Angwin et al. titled "Machine Bias," which argued that COMPAS was biased against African American defendants.[14] ProPublica analyzed a data set containing the identity of thousands of defendants, together with their COMPAS score for risk of recidivism and whether they were actually arrested during the two years that followed the COMPAS assessment.[15]

The key result of the analysis, as well as the cornerstone of the story, was that COMPAS erred differently for black and white defendants. Angwin et al. reported that the false positive rate (i.e., the rate at which defendants were predicted to recidivize, but did not) was 38% for black defendants, compared to 18% for white defendants. Conversely, the false negative rate (i.e., the rate at which defendants were predicted not to recidivize, but did) was 38% for black defendants, compared to 63% for white defendants. In other words, overestimation of risk seemed more likely for black defendants, and underestimation of risk seemed more likely for white defendants. One concern with this result is that COMPAS does not predict recidivism as a binary variable but delivers instead a risk score from 1 to 10. In order to compute false negative and false positive rates, it is necessary to choose an arbitrary cutoff above which COMPAS is considered to predict recidivism. The results of Angwin et al. are based on a cutoff of 5, and some critics argued that this arbitrary choice discredited the main finding of the report.[16] However, a reanalysis of the ProPublica data assuages this concern

by showing that the main finding of the report holds for *any* choice of cutoff (Figure 3.1).

An algorithm whose mistakes are unfair to black defendants clearly raises ethical issues, but does it reflect an ethical *trade-off*? In this specific case, the answer appears to be yes, because two conceptions of fairness can apply whose simultaneous satisfaction is mathematically impossible.[17] In essence, the algorithm can be equally predictive for both groups, or equally wrong for both groups, but not both. The algorithm is equally predictive for both groups when the probability of recidivism is the same for two individuals who have the same score, regardless of their group. The algorithm is equally wrong for both groups when it yields the same rate of false positives and false negatives for both groups. However, and this is the critical point, these two properties cannot be simultaneously satisfied if the two groups do not have the same baseline probability of recidivism. As soon as one group has a greater recidivism rate, one must decide where to put the cursor between equal predictive power and equal mistakes. It is obvious that unequal mistakes are unfair. And yet, if they are equalized, the risk score must be interpreted differently for black and white defendants. A score of 6 could denote a high risk for a white defendant and a low risk for a black defendant—which means that judges using the algorithm would necessarily factor race into their sentences, something that they are currently forbidden to do.

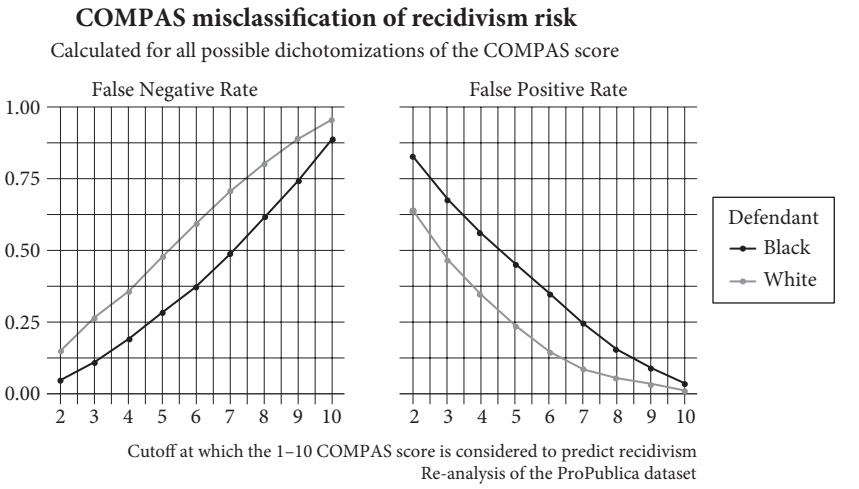


Figure 3.1 A reanalysis of the ProPublica data set shows that the main result of Angwin et al., "Machine Bias," holds for all dichotomizations of the COMPAS score, assuaging concerns that the result was linked to the arbitrary choice of cutoff in the original article.

We do not intend to explore the legal ramifications of such a transformation of judicial practices. Rather, our goal in discussing the COMPAS controversy was to show that some ethical trade-offs will be much harder than others to stylize for laypersons, and thus much harder to study with the standard methods of moral psychology. The problem, though, is that these trade-offs may be the ones most in need of psychological investigation. We can venture that the number of persons who heard about the ProPublica story is orders of magnitude larger than the number of persons who know about the trade-off it reflects—and we can imagine as a result that many people believe that sentencing algorithms are intrinsically racist. If we are to gauge the social acceptability of sentencing algorithms, behavioral scientists will have to uncover an appropriate method to make their ethical trade-offs as understandable as the trolley problems made the ethical trade-offs of AVs understandable to a general audience.

3.3. Second-Order Ethical Challenges

Even if we can develop appropriate methods to measure social preferences and expectations about machine ethics, and even if we perceive the benefits of doing so, we need to be careful about the unintended negative consequences of such experiments. In other words, we need to be mindful of the second-order ethical challenges involved in conducting psychological studies of machine ethics. Here we consider two such concerns: that studies of machine ethics may lead to a waste of resources, and that studies of machine ethics may unduly scare the public. It is important to note right away that these concerns are proportional to the media attention that studies receive, for reasons that will be apparent shortly.

3.3.1. Wasteful Studies

Many speakers who have given talks on AVs to popular audiences have had the same experience: whatever their specific topic was, they got a question about trolley problems. That is, not only did one specific method (trolley problems) capture the attention of the media and the public to the point of becoming synonymous with AV ethics, but it threatened to dominate the conversation on AVs to the detriment of more central questions such as overall safety and environmental efficiency.

The concern that we have often heard is that such a fixation may lead car companies and policymakers to make wasteful decisions. For car companies, a wasteful decision would be to commit too many resources to addressing trolley-like dilemmas (which companies are ill-equipped to deal

with anyway, for the lack of staff trained in ethics) and not enough resources to improving safety and avoiding such dilemmas in the first place. While there may be some theoretical point at which spending on ethics becomes a wasteful extravagance, we argue that we are not yet close to approaching this point. Though we are not privy to the financial decisions made by car companies, the fraction of resources that these companies devote to ethical issues is most likely an infinitesimal portion of the resources that they devote to engineering issues. Being thrifty about any aspect of safety (absolute or relative) would be a suicidal move for an AV company, which suggests that we should not be overly concerned about ethical teams absorbing the resources of engineering teams.

When it comes to our other examples, KPD and sentencing algorithms, the situation is quite different because these algorithms are already in place and already raising ethical questions or concerns. Here it seems that devoting *more* resources to these ethical issues would be a *good* move, especially in the case of sentencing algorithms—even if it means that some resources might be diverted from the technological refinement of the algorithms. Overall it would seem that market forces are more than enough to counter any tendency to overspend on ethics and underspend on performance. Furthermore, the risk of ethical opt-out means that money spent on ethics is not *wasted* since performance without adoption is useless.

Policymakers, though, may find themselves under pressure to act too fast or too strongly in order to assuage the fears of their constituencies, if these constituencies identify ethics as the sole or most pressing issue regarding the use of AI. The antidote, though, is to conduct more psychological studies, not fewer—as long as these studies can appropriately inform policymaking. The faster we can inform policymakers of what citizens are willing or unwilling to accept, the lower the risk that policymakers will make hasty decisions that hamper the development of AI for no good reason.

In sum, the toothpaste is out of the tube now that the general public is aware of the challenges of machine ethics; there is no going back. Psychological studies of machine ethics will not cause wasteful decisions, but the lack of such studies surely will.

3.3.2. Scary Studies

A related but different concern with studies on machine ethics is that we can adversely affect the public attitude toward AI by the process of measuring it. Consider again the focus on trolley problems in studies of AV ethics. Trolley-like situations are very aversive while being (in their literal and simplified form)

extremely rare. Drawing attention to these situations, then, may adversely and irrationally affect the subjective perception of the safety of AVs.

When thinking of small probability events, people are prone to several biases that include the availability heuristic (risks are subjectively higher when they come to mind easily)[18] and the affect heuristic (risks are subjectively higher when they evoke a vivid emotional reaction).[19] Because AV trolley situations can be easily imagined (whatever their actual probability of occurrence), and because they plausibly trigger a strong emotional reaction, the danger is that their likelihood may be overestimated, with downstream consequences on the acceptability of AVs in general. Worse, this impact may be compounded by algorithmic aversion (people lose confidence in erring algorithms more easily than for erring humans).[20] This is an important problem, but once more, it will not be solved by keeping ethical dilemmas out of public sight. In June 2016 the first fatality involving a car in self-driving mode drew more media attention than the approximately fifteen thousand human-driven car accidents that occurred in the United States on that same day. We can only imagine the coverage of the first fatality that will occur when an AV faces something akin to a trolley dilemma. Before it comes, the public should have discussed it openly and had a voice in how the AV was programmed to act, rather than been kept in the dark.

In any case, whether people are deterred by AV trolleys is an empirical question deserving of actual research. To explore this question, we conducted a survey on the Amazon Mechanical Turk platform, recruiting four hundred participants from the United States, of which 369 completed the full survey. Participants were randomly assigned to either a condition in which they were first exposed to three AV trolley dilemmas, and then to four questions about their attitudes toward AVs (the *dilemma first* treatment), or the reverse order, responding first to the four questions about their attitudes, and only then being exposed to the three AV dilemmas (the *control* treatment). In addition, all participants gave information at the end of the survey on their prior exposure to AV dilemmas,[21] their driving habits, their demographics, and their love of technology (7-item scale). The four questions about attitudes were:

1. How excited are you about a future in which autonomous (self-driving) cars are an everyday part of our lives? (7-point scale from 1 = Not at all, to 7 = Very Much)
2. How afraid are you about a future in which autonomous (self-driving) cars are an everyday part of our lives? (7-point scale from 1 = Not at all, to 7 = Very Much, reverse-coded so that higher scores reflect more comfort with AVs)
3. Should they become commercially available by the time you are next purchasing a new car, how likely would you be to choose an autonomous vehicle? (7-point scale from 1 = Not at all likely: I would rather buy a car without self-driving capabilities, to 7 = Extremely likely: I would definitely choose to buy a self-driving car)

Table 3.2 Attitude toward AVs (95% confidence interval) for participants who read about ethical dilemmas first, and for control participants who read about ethical dilemmas after they expressed their attitudes about AVs. This analysis is restricted to participants who had never heard about the ethical dilemmas of AVs before taking the survey.

	Dilemmas first N = 132	**Control N = 132**	*t*	*p*	**Bayes factor**
Excited about AVs	3.4–4.2	3.9–4.5	−1.6	.11	2.2
Will purchase	2.7–3.3	2.8–3.4	−0.4	.65	6.7
Feels safe	3.6–4.3	3.4–4.1	0.1	.92	7.4
Feels no fear	3.3–3.9	3.5–4.1	−0.7	.48	5.8

4. Compared to current human-driven cars, how safe do you expect self- driving cars to be? (7-point scale from 1 = Much less safe, to 7 = Much more safe)

As shown in Table 3.2, reading about the ethical dilemmas of AVs had no discernible impact on any measure of participants' attitude toward AVs. (The analysis was restricted to the 264 participants who had never heard about the dilemmas before taking the survey; the results are even stronger when the analysis is conducted on the full sample.) In particular, reading about ethical dilemmas did not impact participants' perception of their safety and did not impact their willingness to acquire one. A Bayesian analysis[22] showed that the Bayes factors $Pr(H0|D)/Pr(H_1|D)$ ranged from 2.2 to 7.4, offering positive to substantial evidence for the null hypothesis.

Since participants informed us of their level of exposure to the ethical dilemmas of self-driving cars before taking the survey, we could estimate the impact of this prior exposure on their attitude. Prior exposure to the dilemmas was measured on a 5-point scale (no exposure, little exposure, moderate exposure, a lot of exposure, a great deal of exposure). For the purpose of this analysis, we reclassified participants who had no prior exposure to the dilemmas but who read about the dilemmas first in the study as having "a little" exposure. Figure 3.2 shows the effect of prior exposure on participants' attitudes about AVs. Visual inspection does not suggest that prior exposure has any adverse affect; in fact, the trend is positive, suggesting a positive effect of exposure. This trend, though, appears to result from a statistical confound: respondents with a high level of exposure are also the ones with the highest appreciation for technology.[23] Controlling for this variable (as well as demographic variables), the net effect of

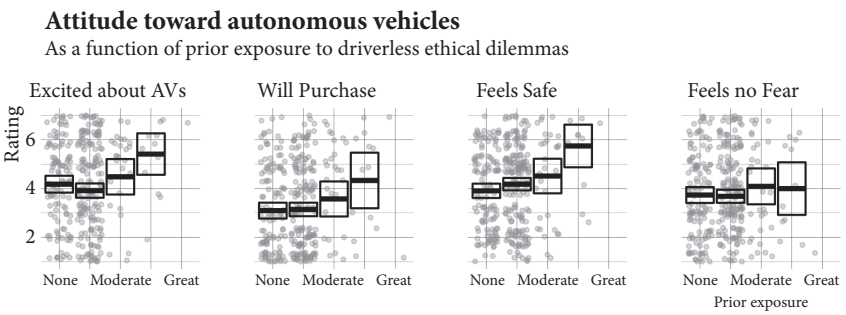


Figure 3.2 Attitude about AVs as a function of the level of prior exposure to the trolley-like dilemmas of AVs. Boxes show the 95% confidence interval of the mean for each level of exposure, except for "great" exposure, for which not enough data points were available.

prior exposure on attitudes is essentially zero, as shown by regression analyses summarized in Table 3.3.

In sum, we did not find any evidence that the mere exposure to trolley-like dilemmas had any adverse impact on attitudes toward AVs, or on their safety in particular. People do not seem to be intrinsically scared by ethical dilemmas, which suggests that we might not have to worry too much about the affect heuristic. They may not like all possible *solutions* to these dilemmas, and they are likely to opt out of buying AVs if the solutions they do not like are implemented[24]—but merely discussing these solutions is unlikely to sow fear and distrust in the public mind. As a result, there is reason to feel comfortable in continuing with experiments and surveys without fear of, as a byproduct, adversely influencing the attitudes they measure.

It is unclear whether we should be concerned that exposing people to the ethical trade-offs embedded in organ transplant algorithms or sentencing algorithms might generate some indiscriminate mistrust of all algorithms in these domains. In the case of sentencing algorithms, the question is probably moot. News media and popular books have already exposed a great many citizens to instances in which these algorithms behaved erratically or unfairly.[25] Exposing study participants and study readers to the *trade-offs* that the algorithms must face is unlikely to lead to any further generalized negativity than has the asymmetric focus on their mistakes or objectionable predictions. In the case of organ transplants, the notion that donors and recipients must be compatible is so deeply rooted in the public mind that it would seem hard for people to object, in general, to any algorithm that would seek to maximize compatibility—even though they may object to other criteria introduced in the optimization

Table 3.3 Attitude toward AVs as a function of prior exposure to their ethical dilemmas, controlling for demographic characteristics. All continuous variables were standardized before analysis.

	Excited	Feels unafraid	Will purchase	Feels safe
Prior Exposure	−0.0004 (0.05)	−0.02 (0.05)	0.04 (0.05)	0.10* (0.05)
Women	−0.19 (0.10)	−0.45*** (0.10)	−0.19* (0.10)	−0.35*** (0.10)
Age	−0.08 (0.05)	0.01 (0.05)	−0.11* (0.05)	−0.06 (0.05)
Usually Driver	−0.51* (0.23)	−0.59* (0.23)	−0.59** (0.22)	−0.60** (0.23)
Usually Passenger 0.02	−0.27 (0.26)	−0.06 (0.27)	−0.40 (0.26)	(0.26)
Old Kids	0.29 (0.18)	0.03 (0.18)	0.33 (0.18)	0.04 (0.18)
Young Kids	0.10 (0.11)	−0.02 (0.11)	0.07 (0.11)	0.01 (0.11)
Income	0.11* (0.05)	0.11* (0.05)	0.09 (0.05)	0.09 (0.05)
Liberals	0.18*** (0.05)	0.15** (0.05)	0.17*** (0.05)	0.20*** (0.05)
Love for Tech	0.35*** (0.05)	0.25*** (0.05)	0.37*** (0.05)	0.27*** (0.05)
Constant	0.47* (0.22)	0.77*** (0.23)	0.55* (0.22)	0.72** (0.22)
Observations R2	369 0.24	369 0.19	369 0.26	369 0.22

Note: *p<.05, **p<.01, ***p<.001

function. Overall it would seem that behavioral scientists are on safe ethical grounds for measuring people's preferences about machine ethics.

3.4. Summary

AI-driven technology is extending to domains where algorithms will make or inform decisions with tremendous consequences on people's lives and

well-being. Machines may decide who survives a traffic accident, who receives a lifesaving organ, or how long one will stay in jail. The promise of AI is to improve on human decisions and save more lives, be it by avoiding accidents, optimizing organ donation chains, or preventing violent crime. But this promise can only come true if people accept that AI may handle the kind of moral trade-offs that were, until now, the reserved grounds of humans. If people are unhappy with the way moral machines are programmed, they can make them irrelevant by opting out of their use. People can refuse to buy self-driving cars, can opt out of being organ donors, and can vote out judges or politicians who allow the use of algorithms in court. To avoid this ethical opt-out, behavioral scientists must give people a voice by using the methods of moral psychology to assess citizens' preferences about the ways machines should handle ethical trade-offs. This is a challenging task, for behavioral scientists will have to find a way to adapt the methods of moral psychology in order to tackle complex technical domains, which are likely to elicit complex moral preferences. Furthermore, behavioral scientists will have to tread carefully and be mindful of second-order ethical challenges. But as we have shown in this chapter, none of these challenges is intractable—and the stakes are great.

Moral psychology has traditionally kept an eye on the past, whether the evolutionary past that shaped our moral intuitions or the work of the great philosophers that formalized ethical theories. It is now time to turn an eye to the future and to investigate the moral psychology of the newly possible.

Notes

1. N. Bostrom, *Superintelligence: Path, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
2. (*Trust in technology*, 2017)
3. J. F. Bonnefon, A. Shariff, and I. Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352 (2016): 1573–76; A. Shariff, J. F. Bonnefon, and I. Rahwan, "Psychological Roadblocks to the Adoption of Self-Driving Vehicles," *Nature Human Behaviour* 1 (2017): 694–96.
4. D. Walton, *Media Argumentation: Dialectic, Persuasion and Rhetoric* (Cambridge: Cambridge University Press, 2007).
5. R. Thaler and C. S. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New Haven, CT: Yale University Press, 2008).
6. I. Rahwan, "Society-in-the-Loop: Programming the Algorithmic Social Contract," *Ethics and Information Technology* 20 (2018): 5–14.
7. J. D. Greene, *Moral Tribes: Emotion, Reason, and the Gap between Us and Them* (New York: Penguin Books, 2014); J. Haidt, "The New Synthesis in Moral Psychology," *Science* 316 (2007): 998–1002.

8. H. Gintis et al., *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (Cambridge, MA: MIT Press, 2005)
9. P. Foot, "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review* 5 (1967): 5–15.
10. N. Goodall, "Ethical Decision Making during Automated Vehicle Crashes," *Transportation Research Record: Journal of the Transportation Research Board* 24 24 (2014): 58–65.
11. Bonnefon, J. F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. Proceedings of the IEEE, 107, 502-504.
12. M. J. Casey et al., "Rethinking the Advantage of Zero-HLA Mismatches in Unrelated Living Donor Kidney Transplantation: Implications on Kidney Paired Donation," *Transplant International* 28 (2015): 401–9.
13. R. Freedman et al., "Adapting a Kidney Exchange Algorithm to Align with Human Values," *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).
14. J. Angwin et al., "Machine Bias," *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
15. A detailed presentation of the analysis is available from Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Data and code can be downloaded from https://github.com/propublica/compas-analysis.
16. A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias," *Federal Probation* 80 (2016): 38–46.
17. A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *arXiv*, February 28, 2017, *arXiv:1703.00056*; J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-offs in the Fair Determination of Risk Scores," *arXiv*, last revised November 17, 2016, arXiv:1609.05807; G. Pleiss et al., "On Fairness and Calibration," In *Advances in Neural Information Processing Systems* (2017), 5684–93.
18. A. Tversky and D. Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology* 5 (1973): 207–32.
19. M. L. Finucane et al., "The Affect Heuristic in Judgments of Risks and Benefits," *Journal of Behavioral Decision Making* 13 (2000): 1–17.
20. B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General* 144 (2015): 114–26.
21. First, participants were asked, "Prior to doing this survey, had you heard any discussion about self-driving cars having to make ethical choices such as deciding who should live and die in an accident?" (yes/no). Participants who responded "yes" were then asked, "You indicated that you had heard about self-driving car ethical issues before. How much thought have you given them?" (5-point scale, from 1 = *None*, to 5 = *A Great Deal*).

22. R. D. Morey and J. N. Rouder, *Bayes Factor: Computation of Bayes Factors for Common Designs* [Computer software manual], R package version 0.9.12–2, May 19, 2018, https://CRAN.R-project.org/package=BayesFactor.
23. For related results, see M. F. Kramer et al., "When Do People Want AI to Make Decisions?," paper presented at the *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, February 13, 2018 at the Hilton New Orleans Riverside, in New Orleans, Louisiana.
24. Bonnefon, Shariff, and Rahwan, "The Social Dilemma of Autonomous Vehicles."
25. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2017).

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. "Machine Bias." *ProPublica*, May 2016.

Bonnefon, J. F., A. Shariff, and I. Rahwan. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (2016): 1573–76.

Bonnefon, J. F., A. Shariff, and I. Rahwan. "The Trolley, the Bull Bar, and Why Engineers Should Care about the Ethics of Autonomous Cars." *Proceedings of the IEEE*. (in press).

Bostrom, N. *Superintelligence: Path, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Casey, M. J., X. Wen, S. Rehman, A. H. Santos, and K. A. Andreoni. "Rethinking the Advantage of Zero-HLA Mismatches in Unrelated Living Donor Kidney Transplantation: Implications on Kidney Paired Donation." *Transplant International* 28 (2015): 401–9.

Chouldechova, A. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *arXiv*, February 28, 2017. arXiv:1703.00056.

Dietvorst, B. J., J. P. Simmons, and C. Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (2015): 114–26.

Finucane, M. L., A. Alhakami, P. Slovic, and S. M. Johnson. "The Affect Heuristic in Judgments of Risks and Benefits." *Journal of Behavioral Decision Making* 13 (2000): 1–17.

Flores, A. W., K. Bechtel, and C. T. Lowenkamp. "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias." *Federal Probation* 80 (2016): 38–46.

Foot, P. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5 (1967): 5–15.

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. Artificial Intelligence, 103261.

Gintis, H., S. Bowles, R. T. Boyd, E. Fehr, et al. *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life.* Cambridge, MA: MIT Press, 2005.

Goodall, N. "Ethical Decision Making during Automated Vehicle Crashes." *Transportation Research Record: Journal of the Transportation Research Board* 24 24 (2014): 58–65.

Greene, J. D. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Books, 2014.

Haidt, J. "The New Synthesis in Moral Psychology." *Science* 316 (2007): 998–1002.

Kleinberg, J., S. Mullainathan, and M. Raghavan. "Inherent Trade-offs in the Fair Determination of Risk Scores." *arXiv*, last revised November 17, 2016. arXiv:1609.05807.

Kramer, M. F., J. S. Borg, V. Conitzer, and W. Sinnott-Armstrong. "When Do People Want AI to Make Decisions?" Paper presented at the *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. 2018.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Morey, R. D., and J. N. Rouder. *Bayes Factor: Computation of Bayes Factors for Common Designs* [Computer software manual]. R package version 0.9.12–2. May 19, 2018. https://CRAN.R-project.org/package=BayesFactor.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books, 2017.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. "On Fairness and Calibration." In *Advances in Neural Information Processing Systems,* 5684–93. 2017.

Rahwan, I. "Society-in-the-Loop: Programming the Algorithmic Social Contract." *Ethics and Information Technology* 20 (2018): 5–14.

Shariff, A., J. F. Bonnefon, and I. Rahwan. "Psychological Roadblocks to the Adoption of Self-Driving Vehicles." *Nature Human Behaviour* 1 (2017): 694–96.

Thaler, R., and C. S. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven, CT: Yale University Press, 2008.

Trust in technology (Tech. Rep.). (2017). HSBC.

Tversky, A., and D. Kahneman. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (1973): 207–32.

Walton, D. *Media Argumentation: Dialectic, Persuasion and Rhetoric*. Cambridge: Cambridge University Press, 2007.

4

Modeling and Reasoning with Preferences and Ethical Priorities in AI Systems

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable

4.1. Introduction

Whatever we do in our everyday life, be it at work or in our personal activities, we need to make decisions: what to eat, where to go on vacation, what car to buy, which route to take to work, what job to choose, and many more. Some of these decisions are made in isolation, without any consultation with other individuals. But many of them are collective decisions that we make together with others. To make all these decisions, we usually rely on our subjective preferences over the possible options. If we need to buy a car, we may have preferences regarding its color, its maker, its engine type, and many other features. If we need to decide which restaurant to go for dinner, we may have preferences of location, facilities, food, drinks, and many other features.

In many domains our subjective preferences are combined with moral values, ethical principles, or behavioral constraints that are applicable to the decision scenario.[1] We have our own preferences over food, but maybe the doctor recommended that we follow a diet to avoid some health issues, so we need to combine the doctor's guidelines with our taste preferences.[2] This is especially true in decisions that may have an impact on others. In many of these contexts, computer science views these interactions as *multiagent systems*.[3] In a multiagent system, the different preferences and priorities, one for each of many agents, are considered at the same time. Typically, work in these fields has to do with modeling,[4] aggregating,[5] and reasoning[6] with these possibly competing agent preferences and priorities. It is important to note here that in this chapter we will discuss a descriptive and measurement-based framework for making decisions, with examples. In this work we do not seek to make normative judgments about what *should be done*. Rather, this chapter shows how methods from computer science and preference reasoning can possibly be used to compare two, possibly conflicting, preferences.

We see the multiagent context as instructive for AI ethics as in many contexts one can view the competing priorities of various agents. Additionally, social

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable, *Modeling and Reasoning with Preferences and Ethical Priorities in AI Systems* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0005

norms, regulations, and laws could provide guidelines to follow when making a decision[7] and also be modeled as yet another agent in the system. While driving our car, we may want to drive as fast as possible to get home sooner, but social norms and laws provide limits to speed and dangerous driving behavior. When trying to inject subjective preferences and ethical priorities in a machine, it is important to be able to model these concepts, reason with them, and combine them, while at the same time to keep them separate in order to give them different weights.[8]

Since decision-making is a central task in AI systems, the study of how to represent,[9] learn,[10] and reason[11] with preferences has been very active within artificial intelligence and beyond,[12] with significant theoretical and practical results[13] as well as open-source libraries and data sets.[14] In many scenarios that include multiagent systems[15] and recommender systems,[16] user preferences play a key role in driving the decisions the system makes. Thus AI researchers have defined preference modeling frameworks that are:

- **Expressive**: they allow for many types of comparisons both numeric and ordinal.
- **Compact**: they do not require a large amount of space to represent.
- **Easy for elicitation**: we can learn the preferences with just a few queries to the user.
- **Explainable**: formal preference models have an explicit representation (typically in logic) that can be used as the basis for justifying the outcome of an automated system.
- **Efficient**: for reasoning and aggregation; that is, they do not require too much computational power manipulate.

A number of compact preference representation languages have been developed in the literature for representing and reasoning with preferences; see the work of Amor et al.[17] for a survey of compact graphical models. In this paper we specifically focus on conditional preference structures (CP-nets)[18] but also mention hard and soft constraints[19] and GAI-nets[20] as they have been used widely across computer science and other disciplines.

Having a formal structure to model preferences, especially one that directly models the possible dependencies between various features and/or options, enhances the transparency and explainability of automated decision-making systems. These structures make explicit the properties of the model; that is, they are "white box" models instead of opaque or "black box" models.[21] These white box models are more able to support downstream reasoning based on inference and causality analysis on identifiable features of the preferences and priorities. Explaining nonexplicit models is an important

and emerging research topic, and many researchers have identified explanation and transparency as the cornerstones of trusted AI systems.[22] For example, the Engineering and Physical Science Research Council's Principles of Robotics requires the implementation of transparency in robotic systems and autonomous decision-making systems as "a mechanism to expose the decision making of the robot."[23] Hence we feel that using explicit preference and priority models is necessary to build trustworthy AI systems, since they can support reasoning based on inference and/or causality and provide mechanisms for explainability.[24]

Subjective preferences may apply to one or more of the individual components of a complex decision rather than to the whole decision. If we are choosing a car, we may prefer certain colors over others, and we may prefer certain makers over others. We may also have conditional preferences, such as in preferring red cars if the car is a convertible. For these scenarios, the CP-net formalism[25] is a convenient and expressive way to model preferences that has been used widely in the preference reasoning community.[26] CP-nets provide an effective, compact way to qualitatively model preferences over decisions (often called outcomes) with a combinatorial structure. CP-nets are also easy to elicit and provide efficient methods for optimization, search, and reasoning.[27] Moreover, in a collective decision-making scenario, several CP-nets can be aggregated, for instance, using voting rules,[28] to find compromises and reach consensus among several decision-makers.

As mentioned, often subjective preferences are not enough to make a decision. We also need to consider ethical priorities or social norms.[29] Depending on the context, we may have to consider specific ethical principles or laws derived from an appropriate ethical theory or local statutes.[30] While subjective preferences are important, when preferences and ethical principles are in conflict, usually the latter should override the subjective preferences of the decision-maker.[31] For example, in a hiring scenario, the preferences of the hiring committee members over the candidates should be measured against ethical guidelines and laws, for example, ensuring gender and minority diversity.

The ability to model and reason with ethical priorities is essential also to build trust in AI systems; to achieve this, we need to provide these systems with the ability to discriminate between what one would broadly call "good" and "bad" decisions according to some moral values.[32] This means that the quality of a decision should be based not only on the preferences or optimization criteria of the decision-makers but also on properties related to the impact of the decision, such as whether or not it is ethical or legal according to constraints or priorities given by any number of exogenous sources. Indeed there may be specific ethical principles, depending on the context, that could and should override the subjective preferences of the decision-maker.[33]

We argue that it is essential to have systematic and rigorous methodologies to evaluate whether preferences are compatible with a set of ethical principles and to measure the difference between the preferences and the ethical principles. In our work we assume we are given a CP-net, which is an efficient way to represent an ordering over all the elements in a domain. For example, a CP-net over all the movies at a theater would tell us something like "I would most prefer to go see a comedy movie; if I cannot, then I next prefer to go see an adventure movie; if I cannot, then next I prefer to go see an action movie." If we were to think about just one CP-net, then we can say things like the distance (or difference) between getting to see a comedy movie and an action movie, for you, is three places in your preference ordering. Our recent technical work has been on comparing not just within one person's preference but between two complete specifications of all the preferences.[34]The ability to precisely quantify the distance between subjective preferences and external priorities, such as those given by ethical priorities, provides a way to both recognize deviations from feasibility or ethical constraints and also to suggest more compliant decisions.[35]

Informally, this formal distance between CP-nets will allow us to measure how different two preferences expressed as CP-nets are from each other. Mathematically, one defines a distance function or metric between two points, A and B, that must satisfy the following four criteria: (1) the distance between A and B must be $\geqslant 0$; (2) the distance between A and B must be the same as the distance between B and A, that is, it must be symmetric; (3) the distance between A and B must be less than or equal to the distance to any other third point C, that is, it must satisfy the triangle inequality; and (4) if the distance between A and B is zero, then $A = B$. In our example about the movies, we used the Spearman footrule distance,[36] which measures the position in the list of the movies you wanted to see versus the one that you did see. We will more formally define our distance function for CP-nets in section 4.5.

In this chapter, we discuss how to use preference modeling formalisms, such as hard constraints, soft constraints, and CP-nets, to model both subjective preferences and exogenous priorities, such as those provided by ethical principles.[37] We also show how to define and use a notion of distance between CP-nets that can be used to evaluate the distance between an individual's subjective preferences and its exogenous ethical principles, or the ethical principles of a community.[38] We then show how to use this notion of distance between CP-nets to evaluate and guide the decisions made by autonomous AI systems.[39]

With the ability to model both preferences and ethical priorities, as well as to measure the distance between different preferences and priorities structures, we provide a (contextual) solution to the so-called value-alignment problem, which is concerned with being able to model values and check alignment, compatibility, or distance between different values.[40] In our context, values are modeled

as priorities over decisions, induced by either subjective preferences or ethical priorities, or their combination, and we say that two values are aligned if there is some distance that is *small enough* between the two priorities' structures. Note that what "small enough" means will vary by context, but that since we use a distance metric, if this value is set to zero, then the priorities must be *exactly* the same. Specifically in this chapter, we will discuss this when the priority orderings are induced by two CP-nets.

Since CP-nets are a compact representation of a partial order over the possible decisions, the ideal notion of distance is a distance between the induced partial orders of the CP-nets. However, the size of the induced orders may be exponential in the size of the CP-net, and in general computing a distance between these induced partial orders is computationally intractable. Therefore we use a tractable approximation, called *I*-CPD by Loreggia et al.,[41] that is computed directly over the CP-net dependency graphs and has been shown to exhibit a limited error with regard to the correct distance. We detail the value alignment procedure first discussed by Loreggia et al.[42] that computes the distance between subjective preferences and ethical principles and makes decisions using the subjective preferences only if they are *close enough* to the ethical principles, where being *close enough* depends on a threshold over CP-net distances. These thresholds are context-dependent and will need to be decided upon by all the stakeholders of the system—the designers, implementers, community, and leaders—where they will be deployed. If instead the preferences diverge too much from the ethical principles, we move to a less preferred decision until we find one that is a *satisfactory compromise* between the ethical principles and the user preferences. The compromise is defined by setting a second threshold over distances between decisions.

Compactness in modeling both preferences and moral values is a necessity when it comes to implementations for artificial agents.[43] Humans are very good at abstracting away details that are not relevant for decision-making and perceive as atomic even complex events or objects that would require a large amount of details to be formally described. Artificial agents don't have this luxury. They rely on combinatorial structures for the vast majority of the knowledge they acquire and store. This is true also when it comes to preferences. A key challenge that has been tackled by the area of knowledge representation has been that of mapping orderings over large sets of options into compact (graphical) models while trying to minimize the information that is lost in doing so.[44]

Since ethical principles define the same kind of structures as preferences, that is, priority orderings over the possible decisions,[45] it is reasonable to conjecture that ethical requirements will also need to be modeled compactly in order to be embedded into a machine. In preferences research, a compact model is one that can be written down in a smaller number of bits than the overall

preference. Compact models are important because they are space-efficient and do not require us to write down, for example, all pairwise comparisons between a large set of objects. One may argue that there are alternatives available. For example, one could take a machine-learning–based approach where "ethics" is modeled by one or more learning modules trained on, for example, dilemmas and corresponding solutions.[46] While this approach may be feasible, it does raise some concerns. For example, it may not be acceptable that the artificial agent will not be able to provide an explanation for why it judged one action "more ethical" than another. Moreover, as noted in many papers in the literature,[47] bottom-up approaches to ethics tie the results to the data on which the module is trained. This may lead to undesirable outcomes if the data are biased or not general enough. In this paper we take a top-down approach; we assume that the preferences and priorities are articulated through a given CP-net. However, in the future one could imagine this CP-net being learned from data, giving us a more bottoms-up flavor.[48]

4.2. Background: Frameworks to Model Constraints and Preferences

As we have seen, modeling preferences is a topic that has received great attention in the computer science literature. Here we discuss a few of these formalisms with an eye to using them to model both preferences and ethical priorities.[49]

4.2.1. Hard and Soft Constraints

Hard constraints,[50] usually just called constraints, model restrictions on the combination of values that some decision variable can take. For example, in a scenario where we need to schedule activities over time, we may use one decision variable for each activity, which can take values from the time line, and we may pose the constraint that activity A has to occur before activity B. Thus, with a hard constraint, each combination of values of variables A and B is either feasible, that is, it satisfies the constraint, or not. Given several such constraints, the global scenarios that are declared feasible are those in which all constraints are satisfied.

Soft constraints generalize the notion of constraints to allow for more than just two states (feasible or not) for the value combinations of activities. More precisely, a *soft constraint* involves a set of decision variables and associates a value from a (totally or partially ordered) set to each instantiation of its variables.[51]

This allows for a more fine-grained notion of constraints, thus resulting in an ordering.

For example, in the activity-scheduling example described earlier, we may work with a preference structure that includes totally ordered values between 0 and 1, where a higher value denotes a higher preference, and we may have a soft constraint assigning value 0 to combinations of values (A = a, B = b) where $a \not\leq b$, and value $(b-a)/b$ to the other combinations of values, meaning that we do not allow A to occur after B, and when A is before B, we prefer these two activities to be as close as possible. It is easy to see that hard constraints are just a specific class of soft constraints where we have just two preference values, for example, true and false rather than many, and they are combined via logical conditions; that is, all constraints must be satisfied to have a feasible decision.

Fuzzy Constraint Satisfaction Problems (CSPs)[52] are another specific class of soft constraints where we work with values between 0 and 1. In this case higher values denote higher preference and we combine the values with the *min* operator; that is, we take the worst preference, which by definition is the min of the two values. In other words, in fuzzy constraints the goal is to maximize the minimum preference. Because of their nature, fuzzy CSPs are useful when we have safety-critical applications such as self-driving cars or medical devices, since we focus on the worst preference value when we evaluate a complex decision.

Yet another generalization is known as Weighted CSPs. In a Weighted CSP the preference structure contains natural or real values, interpreted as costs or penalties, meaning that a lower value denotes a higher preference, and preferences are combined by summing the values. In other words, weighted constraints aim to minimize the sum of the costs. In computer science, reasoning and decision problems are often judged by their computational complexity. Complexity is a worst-case measure of how hard a problem is to answer in the limit. Generally speaking, problems that require only polynomial time in the size of the input are considered "easy" problems, while those requiring more time or even nondeterministic time are "hard."[53] In general, finding an optimal solution for a hard or a soft constraint set is computationally hard.[54] However, it is polynomial for some classes of (soft) constraints.[55] In this chapter we will see that measuring the distance between CP-nets is computationally hard, but that we have developed efficient and accurate approximation algorithms that work well in practice.[56]

4.2.2. CP-nets

CP-nets, short for Conditional Preference networks, were first proposed by Boutilier et al.[57] They are a graphical model for compactly representing

conditional and qualitative preference relations. For readers who are familiar with Bayesian networks,[58] CP-nets resemble Bayesian networks but replace probabilities with preferences.

CP-nets consist of sets of ceteris paribus preference statements (cp-statements).[59] For instance, the cp-statement *I prefer red wine to white wine if meat is served* asserts that, given two meals that differ *only* in the kind of wine served *and* both containing meat, the meal with red wine is preferable to the meal with white wine. Note that in general one could define cyclic CP-nets where, for example, the choice of wine is dependent on the choice of main, which is dependent on the choice of wine. However, it is both hard to interpret the semantics of cyclic CP-nets and computationally hard to reason with them as well.[60] Indeed it would be hard to understand humans with cyclic preferences too. Due to these issues, we restrict our discussion to *acyclic* CP-nets only.

Formally, a CP-net has a set of features. In the earlier cp-statement, the features involved are the *type of wine* and the *meal*. For each feature, we are given a set of *parent* features that can affect the preferences over its values. In this example, feature *type of wine* depends on *meal*, which is the parent feature. This defines a *dependency graph* in which each node represents a feature and has its parent features as its immediate predecessors. An *acyclic* CP-net is one in which the dependency graph is acyclic. Given this structural information, one needs to specify the preference over the values of each variable for *each complete assignment* to the parent variables. This preference is assumed to take the form of a total order over the values of the dependent variable(s).

The semantics of CP-nets depends on the notion of a worsening flip. A *worsening flip* is a change in the value of a variable to a less preferred value according to the cp-statement for that variable. This definition induces a preorder over the outcomes, which is always a partial order if the CP-net is acyclic. Finding the optimal outcome of a CP-net is computationally difficult (NP-hard).[61] However, in acyclic CP-nets, there is only one optimal outcome, and this can be found in linear time by sweeping through the CP-net dependency graph while following the dependencies, assigning to each feature the most preferred values in the corresponding cp-table. Much of the research in CP-nets over the years has investigated restrictions and testing for dominance and consistency between preferences.[62]

4.3. Modeling Ethical Theories via Hard and Soft Constraints

Many ethical theories have been defined and are used to modeling human behavior when deciding what actions to take. We provide a brief

summary here but refer the reader to more extensive resources for complete definitions.[63]

- Utilitarianism: Action consequences are evaluated on a numerical value scale from good and bad, and an agent should choose the action that maximizes the net expected value of its actions; that is, it chooses the action that has the greatest difference between the benefits and harms that result. This is not the same as choosing the action with the greatest benefit nor the action with the least resulting harm.
- Virtue Ethics: An agent should choose actions that satisfy some predefined set of virtues.
- Deontology: Actions are predefined as right or wrong (via, e.g., the categorical imperative), and an agent should choose the right action, no matter the consequences.[64]

Hard constraints appear to be ideal for modeling deontological ethics. One could envision defining constraint problems where the actions under consideration are complete assignments to a set of decision variables modeling their different aspects and components. The constraints would be modeling ethical restrictions. Then an action would be defined as permissible if it is one of several solutions to the constraint problem, as impermissible if it is not a solution, and as obligatory if it is the only solution.

Soft constraints also have many appealing properties in terms of what may be desired for modeling ethical requirements. First of all, any partial order can be represented. This is not true for other models, such as, for example, CP-nets. This is important in this context because ruling out some orderings may mean that the model may be able to represent not the "true" ethical ordering but only an approximation. Another interesting feature of soft constraints is that different combinations of operators can be chosen in order to aggregate preferences from different constraints. This can be useful if we want to model different ethical theories.

Weighted constraints appear the natural choice when it comes to modeling utilitarianism, which aims at maximizing utilities. In fact it easy to translate the principle of maximizing utilities to that of minimizing costs. On the other hand, fuzzy preferences, which are aggregated with the min operator, well represent the fact that a violation of "ethical" constraints on any component should affect the quality of the entire option. The fundamental question is this: What is the set of properties of a preference aggregator that makes it suitable for handling ethical requirements? Some may be obvious, for example, commutativity. Others may be a point of discussion, such as, for example, the fact that the aggregation of

two ethical preferences cannot be "more ethical" than each of the two original preferences.

4.4. Using CP-nets to Model Preferences and Ethical Priorities

While soft constraints' quantitative approach to preferences may be appealing for modeling some theories, there are other theories that cannot be easily quantified. Some are what MacAskill[65] calls "ordinal theories." Under ordinal theories there is no explicit numerical value associated with options, but rather an ordering. For such theories qualitative preferences, such as those modeled in CP-nets, may be a better option.

Several properties of CP-nets look appealing for the objective we consider. First of all, being able to model conditional statements may be desirable. While one may argue that ethical principles should be absolute and not context-dependent, the study of several dilemmas, such as the Trolley Problem,[66] have shown that what humans regard as ethical may very well be dependent on the context and sometimes for not a very clear or rational reason. CP-nets also have the quality of not requiring numbers to express preferences. It has been argued that numbers may be a cumbersome and tedious way of representing even mundane preferences. When it comes to ethical requirements this argument may become even stronger.

One issue concerning CP-nets is that, as mentioned, some orderings may not be represented. Furthermore, given two decisions, understanding whether one is more desirable has a very high computational complexity. This may be unacceptable in situations where the agent is confronted by a dilemma involving two options, both with some catastrophic effect, and a decision must be made in a short amount of time.[67]

We will now describe with an example how CP-nets can be used to model ethical principles that may come from one or more ethical theories or societal value systems.[68] Our example models a scenario wherein autonomous or human-operated vehicles operate.[69] Each driver (or vehicle) has his or her own subjective preferences or ethical priorities over the possible actions to take in situations for which the traffic laws do not prescribe a specific behavior. Moreover there can be collective ethical guidelines that a community may come up with and would like all drivers to follow, with some tolerance. We claim there are many scenarios wherein subjective preferences and ethical priorities should be considered, combined, and possibly compared. For example, again in the autonomous vehicle setting, the Moral Machines project[70] collected a large number of pairwise choices and analyzed the data from various points of view. Just as in that project, we take

inspiration from the Trolley Problem to design our example. For ease of comparison, we discuss the same example as used in our other work where we evaluate the CP-net distance metrics empirically and discuss their use.[71] An interesting direction for future work is extending both the ethical and moral theories as well as the preferences and priorities to work with *probabilities*. In real life there are probabilities associated with our actions, and understanding how to define and reason with probabilistic structures is an interesting problem. There exist probabilistic versions of CP-nets already,[72] and understanding how to combine them with ethical reasoning is an important next step.

Suppose a vehicle has a brake malfunction while approaching an intersection where cats and humans are crossing the street. The driver has two options: go straight or swerve. If the driver goes straight, she then has the unfortunate option of running over either cats or humans, injuring some of them but saving all of the passengers in the vehicle. On the other hand, the driver can swerve off the road, which will result in saving both the cats and the humans but injuring all the car passengers. This setting is modeled by a CP-net with three features: the kind of hurt entities, the number of victims, and whether the car (or the human driver) decides to go straight or to swerve. Each feature has two possible values, so there are in total eight possible scenarios. Preferences are on each feature, in the form of a total order between the two values of the feature.

In what follows we suppose that the ethics of the various drivers and those of society come from one of any number of sources: laws, standards, best practices, local customs, or any another source or stakeholder. Again, we are not making normative judgments about what communities *should* enact. Rather we are giving a description of a framework by which they can *reason* about how agents act. For example, if a community decides to be "mean" in the following examples, then that is the will of the community and the framework can support decision-making around a defined set of priorities. What follows is just an example; at implementation and deployment time there can and must be a broad, multi-stakeholder conversation about what are the correct morals for these systems.[73]

Figure 4.1 shows the preference of a diligent driver (called Driver 1), modeled by the CP-net shown in the left part of the figure. In this case, this driver prefers to hurt cats rather than people, as can be seen in the preferences over the values of the upper feature. Also she prefers to hurt as few people (or cats) as possible, and she prefers to go straight. The three features are independent of each other, meaning that the preferences over each feature do not depend on the value of any other feature.

The right part of Figure 4.1 shows the ordering over the eight scenarios induced by the CP-net, according to its preferences. This partial order has eight elements, and the directed arcs denote preference dominance. According to this driver's preferences, the optimal scenario (at the top of the partial order) is one

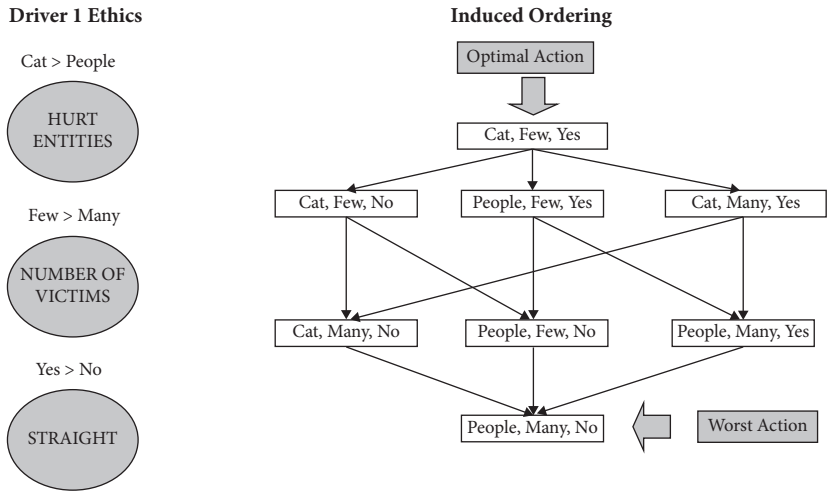


Figure 4.1 On the left side: a CP-net that models the moral preferences of Driver 1. On the right side: the induced partial order over the action space.

where she goes straight, injuring few cats (and saving the human pedestrians). It is possible to see that there are also pairs of scenarios that are not comparable, such as $\{cat, few, no\}$ and $\{people, few, yes\}$, meaning that the driver's preferences do not allow her to establish a priority over them.[74]

In Figure 4.2 we show the preference of a nasty driver; let us call him Driver 2. Although we hope no such drivers exist, for the sake of this example let us assume they do exist and that they prefer to run over the greatest number of humans. For the sake of the comparison, in the induced partial orders of Figure 4.2 and Figure 4.1 we used the same outcomes position but rearranged the arrows. It is easy to see that his CP-net has the same features and values as that of Driver 1, but very different preferences. This is reflected in the ordering induced over the eight scenarios, where the optimal action is now to injure many people by going straight.

Let us now model the ethical priorities of the community in which these two drivers will act. Figure 4.3 describes the moral preferences over the possible solutions that we assume are derived from some appropriate ethical theory or have been decided upon through a collective effort in a society.[75] For our example, we assume the community prefers not to kill human pedestrians and save as many lives as possible. In cases where killing someone is unavoidable, then we assume it is morally preferable to have the smallest number of victims. Finally, we assume it is acceptable to possibly cause injuries to the passengers when there is a sufficiently large number of human pedestrians in danger in the street. To

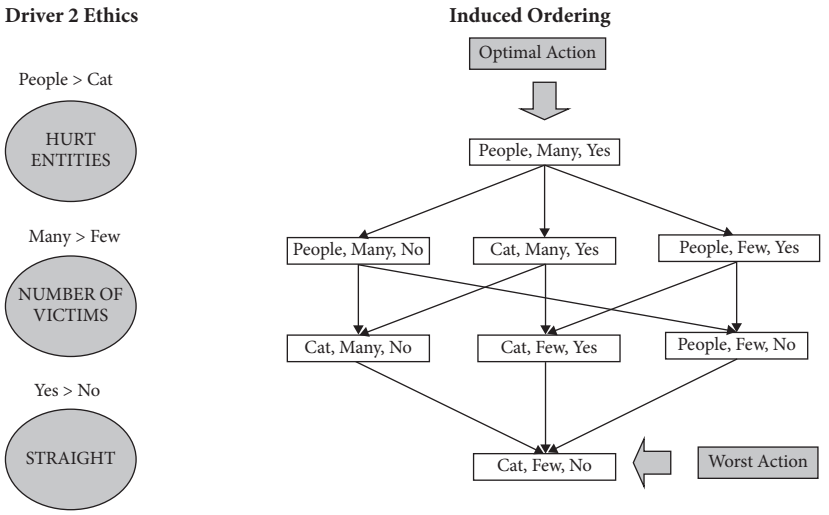


Figure 4.2 On the left side: a CP-net that models the moral preferences of Driver 2. On the right side: the induced partial order over the action space.

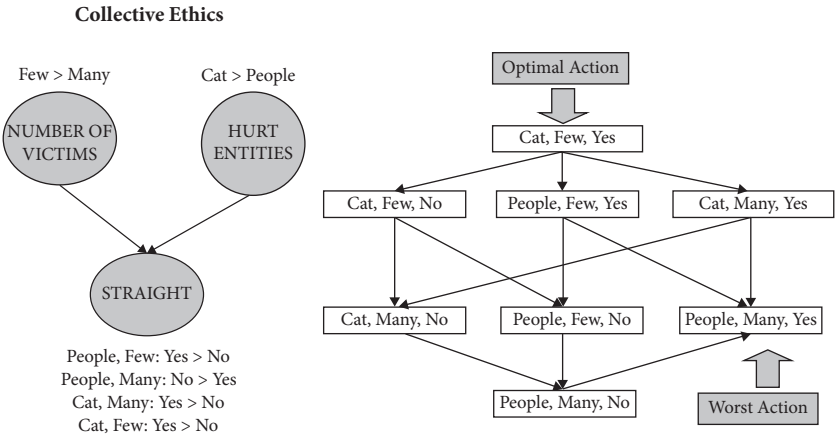


Figure 4.3 On the left side: a CP-net that models the ethical priorities of the community. On the right side: the induced partial order over the action space. Note that we have kept the outcomes in the same positions as in Figure 4.1 and rearranged the arrows, which show the ordering according to worsening flips.

model this, we need to make the direction feature dependent on the other two features, as shown graphically in the CP-net. Thus the preferences over the two values of this feature *Straight* depend on what values are given to the other two features, *NumberofVictims* and *HurtEntities*.

So now we have three sets of ethical priorities, two from the drivers and one from the community. The three partial orders modeling their preferences over possible scenarios are different, since they are induced by different preference structures. How can we understand if a driver is compliant with the community directives over these scenarios? And if not compliant, and the community accepts some tolerance over deviations from the community directives, can we measure how far over he is, in order to understand if he can act according to his priorities or should rather be alerted and guided to act differently?

To answer these questions, it is necessary to have tools that are able to understand how different or how similar two given preference orderings are. This is needed whether the preferences represent moral principles, exogenous priorities, or individual preferences. Having such metrics allows us to define value alignment procedures that enable artificial agents to both follow established guidelines and correct or prevent harm from noncompliant agents. In the next section we discuss these metrics and define a value alignment procedure based on the metrics for supporting decision-making under exogenous priorities.

4.5. A Notion of Distance between (Orderings Induced by) CP-nets

As noted, CP-nets in general do not induce a total order over the possible scenarios but rather a partial order, meaning that some decisions can be incomparable, as seen in the examples in the previous section. A standard notion of distance between two total orders is the Kendall τ distance,[76] which is the number of swaps between adjacent scenarios needed in order to change one order into the other one. For partial orders, we may use something similar, where, however, we must take incomparability into account.

More precisely, we use an extension of the Kendall τ distance with a penalty parameter p defined for partial rankings by Fagin et al.[77] we call KTD. A more complete treatment and proofs of correctness for KTD are discussed by Loreggia et al.[78] Formally, given two CP-nets A and B inducing partial orders P and Q over the same set of outcomes U, we define the Kendall τ distance between P and Q, $KT(P,Q)$, as:

$$KT(P,Q) = \sum_{\forall i,j \in U, i \neq j} K_{i,j}^{p}(P\ Q) \tag{1}$$

where i and j are two scenarios with $i \neq j$, and we have:

1. $K^p_{i,j}(P,Q) = 0$ if i, j are ordered in the same way or they are incomparable in both P and Q;
2. $K^p_{i,j}(P,Q) = 1$ if i, j are ordered inversely in P and Q;
3. $K^p_{i,j}(P,Q) = p,\ 0.5 \leq p < 1$ if i, j are ordered in P(resp. Q) and incomparable in Q(resp. P).

Hence, each pair of outcomes gives a contribution to the overall distance. Given two partial orders, we check each pair of outcomes. If they are ordered in the same way, then their contribution is 0 to the total distance; if they are ordered in opposite ways, their contribution is 1; and if they are ordered and the other one is incomparable, then the contribution is p, which is between 0.5 and 1. Summing all these contributions gives the overall distance between the two partial orders.

In our running example, the distance between the two drivers is 15 (normalized is 0.625); the distance between Driver 1 and the community ordering is 2.5 (0.1136), and that between Driver 2 and the community ordering is 17.5 (0.7291). The distance can be reported either as a total count or as a normalized value. While the unnormalized counts are more intuitive, the normalized values, which are divided by the total number of pairs in a setting, are useful for our later value alignment procedure.

It is easy to see that Driver 1 is much closer to the community ethics than Driver 2. This is reflected in the distance, which is 2.5 for Driver 1 and 17.5 for Driver 2. Also, the distance between the two drivers is 15, which models the fact that they behave very differently.

While this is a very reasonable notion of distance, used effectively in many scenarios, unfortunately computing it is computationally intractable (NP-hard).[79] One could be tempted to forget about the partial orders and define a notion of distance based only on CP-nets. However, small differences in CP-nets may result in huge differences in their induced partial orders. So a correct notion of distance has to be defined over the induced orders. However, we can exploit the information given by the CP-nets that induced such orders to obtain good approximations of KTD. Elsewhere[80] we use this approach to define an approximation, called *I*-CPD, that has a limited error compared to KTD and can be computed in polynomial time in the size of the CP-nets.

4.6. Using Distance to Support Ethical Decisions

We now discuss how to use the distance described earlier to define a value alignment procedure that can alert agents that are not compliant and guide them

toward more ethical actions. This procedure and experiments on a variety of domains we first explored elsewhere;[81] we extend an example here to the domain of autonomous vehicles.

Suppose that ethical principles are modeled via a CP-net N_e and an individual (human or computer) agent models her preferences or ethical priorities via another CP-net N_p. We assume that N_e and N_p have the same set of features with the same values. However, they can differ on the dependency structure, the cp-tables, and therefore the induced ordering, as shown in the running example.

Given the ethical principles and the agent preferences, we need to guide the agent in making decisions that are morally acceptable according to the given ethical principles described by N_e. To do this, we propose the following value alignment procedure:

1. We set two distance thresholds: one between CP-nets, called t_1, that ranges between 0 and 1, and another between actions, called t_2, that ranges between 1 and n, the number of features of both N_e and N_p.
2. We check if the distance between CP-nets N_e and N_p is less than t_1. For example, here we can use the *I*-CPD distance to compute it in a tractable way.
3. If the distance is below the threshold t_1, the agent is allowed to choose the top action in the partial order induced by the CP-net.
4. If the distance is above the threshold t_1, then the agent must move down his induced preference ordering, through worsening flips, to a less preferred action, until he finds an action that is closer than t_2 to the optimal action according to the ethical CP-net N_e. This is a compromise decision between what his preferences say and what the ethical principles recommend.

The fundamental idea in this value alignment procedure is that agents can behave as they prefer *only* when their preferences are close enough[82] to the specified ethical principles. Otherwise the agent must compromise by finding a solution that is closer to the one suggested by the ethics. The procedure depends on the two tolerance values t_1 and t_2, that allow for some deviation between the prescribed ethical priorities and that of the individual agent. If no tolerance is allowed, it is enough to set $t_1 = 0$ and $t_2 = 1$. Intuitively, t_1 lets us control the total distance between the orderings, while t_2 controls the number of variables that need to be changed in order to find a compliant decision. It is important to note that t_2 is the number of worsening flips from the most preferred outcome of the community CP-net. This means that whatever CP-net an unaligned driver may have, if we set $t_2 = 0$ then we are only allowing that agent to act in accordance with the top outcome of the community.

Given the definition of the distance between CP-nets,[83] it is important to observe that if we force the CP-nets to be close together, the top elements will not be too far apart. This is due to the metric properties of the distance we are using: the CP-net that is the farthest away is the one with all preferences reversed, while the closest CP-net is the identical one. Hence it is not the case, due to the structure of how CP-nets are constructed, that the two most preferred elements will be very far apart.

Observe that t_2 is a distance over the individual actions and represents how far from the community priorities we allow "unaligned" people to behave, while t_1 is a distance over the CP-net itself and tells us how far away an agent's beliefs must be before we call it "unaligned."[84] However, this raises an interesting challenge for this approach as an agent who sets her personal preferences far away from the ethical principles may be permitted to do things that an agent who is more in alignment with the principles is not. In general this should not be the case given the earlier observation about the properties of the distance, but more empirical experimentation could shed light on this issue.

Turning to our examples in section 4.4, let us use the no tolerance threshold, that is, $t_1 = 0$ and $t_2 = 1$. Driver 1 can act according to her preferences, since his optimal scenario $\{Cat, Few, Yes\}$ coincides with the optimal scenario of the community CP-net. However, Driver 2 cannot act according to his preferences because they are far from the ethics; instead he has to find a compromise by looking to his partial order and finding an action that is just one flip from the best outcome of the community ordering, this because $t_2 = 1$. Let's look for Driver 2's compromise by first looking at all actions that are one flip from the best outcome of Driver 2; then, if no actions are found that are acceptable for the value alignment procedure, move down one more flip in the partial order of Driver 2, and so on, until an action is found that is closer than t_2 flips from the best outcome of the community ordering. So the first step is looking at actions in the partial ordering of Driver 2 that are one flip away from his best scenario: $\{People, Many, No\}$ cannot be taken because it is not one flip away from the best outcome of the community ordering $\{Cat, Few, Yes\}$, but it is three flips away. Either one of the other two outcomes $\{Cat, Many, Yes\}$ and $\{People, Few, Yes\}$ can be chosen because both are one flip away from the best outcome of the community ordering. The value alignment procedure stops proposing one of the two available actions to the driver.

4.7. Distance and Metapreference

A notion of distance is useful not just in measuring deviations but also to deal with dynamically changing preferences and ethical priorities. In a social context,

individual preferences are transformed over time by incorporating elements from interactions with other members of the group. This is often called "reconciliation" of individual preferences with social reason and takes place in the context of collective choice.[85] To be able to describe the dynamic moving from one preference ordering to the next one (over time), and to make sure that the latter preference orderings are indeed better in terms of morality, one needs to have a way to judge preferences according to some notion of good and bad. Indeed Sen[86] claims that morality requires judgment among preferences. To account for this, he introduced the notion of metaranking, that is, preferences over preferences, which enables us to formalize individual preference modifications. A moral code could then be defined as ranking of preference rankings. That is, the moral code is defined by a structure that, by employing notions such as distance, is able to rank preferences according to their morality level. The distance intrinsic in the moral code can then be useful in measuring the deviation of any social or individual action from the moral code itself.

This approach to morality is appealing from a computational point of view. If we intend to use compact preferences models we must address two key points regarding compactly represented preferences, namely, (1) how to dynamically change them and (2) how to define a notion of distance among them. The first challenge has been partially addressed in the literature. Indeed changing preferences can be seen as a form of preference elicitation or learning. This has been shown to pose some computational challenges for CP-nets[87] and has only partially been studied in the case of soft constraints.[88] The task of dynamically updating has also been studied in CP-nets.[89]

Another possibility is seeing learning moral preferences as resolving uncertainty concerning what is moral. This could be represented, for example, by an extension of CP-nets called PCP-nets, where preferences are expressed by a probability distribution over ordering rather than by a single ordering.[90] Then learning can be modeled as a change in the probability distribution that leads to one in which there is no uncertainty, that is, where one ordering has probability 1.

The second challenge is to define distances over compact preference structures. The metarankings defined by Sen as orderings of orderings would be, in our case, orderings over CP-nets, where the ordering would be induced by the distance of the CP-nets from a reference "moral CP-net." This paper and our "On the Distance between CP-nets"[91] give a solution to this challenge by providing a tractable notion of distance between CP-nets. As far as we know, a distance on soft constraints has not been formally defined. Due to its quantitative nature, one point to clarify is whether the actual values of the preference structure should matter or whether only the relative ordering should count.

A final important direction for our work is to understand how to generalize the decision-making process to include multiple ethical CP-nets. Many of us in

our daily lives attempt to simultaneously satisfy norms and values placed on us by society, our families, our religion, and our work. Understanding how to work with multiple and possibly conflicting priorities is a large challenge.

4.8. Summary

It is imperative that we build intelligent systems that behave morally. For them to work and live with us, we need to trust such systems, and this requires that we are *reasonably* sure they behave according to values that are aligned to human values. Otherwise we would not let a robot take care of our elderly people or our kids, nor a car drive for us, nor would we listen to a decision support system in any healthcare scenario. Of course this is less crucial when the application domain does not include critical situations, like suggesting a friend on social media or a movie in an online selling system. But when the AI system is helping (or replacing) humans in critical domains such as healthcare, then we need to have a guarantee that nothing morally wrong will be done.

We argue that existing preference modeling and reasoning frameworks, such as hard constraints, soft constraints, and CP-nets, can be used also to model and reason with ethical principles and moral codes. This provides a single framework where subjective preferences and ethical priorities can be modeled, combined, and compared. We showed how to model ethical priorities via CP-nets with a running example, focusing on the issue of comparing two priority structures in order to measure the possible ethical deviation (between two individual agents or an agent and the community ethical guidelines). A well-defined notion of distance between ethical priorities is essential also for capturing the dynamic evolution of ethical principles. In our work, we use a tractable notion of distance, and we exploit it to define a value alignment procedure that checks if the preferences of an agent are *close enough* to the ethical principles of his community. If the deviation goes beyond the available tolerance, the agent is guided toward less preferred actions that are, however, compliant with the ethical principles of his community.

Notes

1. F. Rossi, "Moral Preferences," in *Proceedings of the 10th Workshop on Advances in Preference Handling at the International Joint Conference on Artificial Intelligence*, New York, New York, USA, 2016; Joshua Greene, "The Cognitive Neuroscience of Moral Judgment and Decision Making," in *The Cognitive Neurosciences V,* ed. M. S. Cazzaniga (Cambridge, MA: MIT Press, 2014).

2. Avinash Balakrishnan et al., "Using Contextual Bandits with Behavioral Constraints for Constrained Online Movie Recommendation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Editor Jérôme Lang, ijcai.org, 5802 – 5804, 2018; Avinash Balakrishnan et al., "Incorporating Behavioral Constraints in Online AI Systems," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, AAAI Press, 3 - 11* 2019.
3. Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations* (Cambridge: Cambridge University Press, 2008).
4. F. Rossi, K. B. Venable, and T. Walsh, *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice* (Morgan and Claypool, 2011).
5. C. Cornelio et al., "Reasoning with PCP-nets in a Multi-Agent Context," in *Proceedings of the 14th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Istanbul, Turkey, Editors: Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind, ACM Press, 969 – 977. 2015.
6. Peter Stone, "Learning and Multiagent Reasoning for Autonomous Agents," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyperbad, India, Editors: Manuela M. Veloso, IJCAI Press, 12–30, 2007.
7. A. Sen, *Choice, Ordering, and Morality* (Blackwell, 1974); Judith Jarvis Thomson, "The Trolley Problem," *Yale Law Journal* 94, no. 6 (1985): 1395–415.
8. Joshua Greene et al., "Embedding Ethical Principles in Collective Decision Support Systems," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, Editors: Dale Schuurmans and Michael P. Wellman, AAAI Press, 4147 – 4151. 2016.
9. Rossi, Venable, and Walsh, *A Short Introduction to Preferences*.
10. J. Fürnkranz and E. Hüllermeier, *Preference Learning* (Springer, 2010).
11. C. Domshlak et al., "Preferences in AI: An Overview," *Artificial Intelligence* 175, no. 7 (2011): 1037–52; G. Pigozzi, A. Tsoukiàs, and P. Viappiani, "Preferences in Artificial Intelligence," *Annals of Mathematics and Artificial Intelligence* 77 (2015): 361–401.
12. J. Goldsmith and U. Junker, "Preference Handling for Artificial Intelligence," *AI Magazine* 29, no 4 (2009).
13. Domshlak et al., "Preferences in AI"; Pigozzi, Tsoukiàs, and Viappiani, "Preferences in Artificial Intelligence."
14. N. Mattei and T. Walsh, *PrefLib: A Library for Preferences*, accessed February 21, 2020, http://www.preflib.org; N. Mattei and T. Walsh, "A PrefLib.Org Retrospective: Lessons Learned and New Directions," In *Trends in Computational Social Choice*, ed. U. Endriss, 289–309. (AI Access Foundation, 2017).
15. Shoham and Leyton-Brown, *Multiagent Systems.*
16. F. Ricci et al., eds., *Recommender Systems Handbook* (Springer, 2011).
17. Amor, Nahla Ben, Didier Dubois, Hela Gouider, and Henri Prade. "Graphical Models for Preference Representation: An Overview." In *Proceedings of the 10th International Scalable Uncertainty Management*, Nice, France, Editors Steven Schockaert, Pierre Senellart, Springer LNCS 9858 96–111. 2016.
18. C. Boutilier et al., "CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements," *Journal of Artificial Intelligence Research* 21 (2004): 135–91.

19. Bistarelli, Stefano, Ugo Montanari, and Francesca. Rossi. "Semiring-Based Constraint Satisfaction and Optimization." *Journal of the ACM*, Volume 44, Number 2, 201 – 236, 1997; Rossi, Venable, and Walsh, *A Short Introduction to Preferences*.
20. C. Gonzales and P. Perny, "GAI Networks for Utility Elicitation," in *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning,* Whistler, Canada, Editors: Didier Dubois, Christopher A. Welty, Mary-Anne Williams, AAAI Press, 224 – 234. (2004).
21. One could observe that "clear box" may be a better metaphor, but "white box" is fairly standard in the literature on machine learning and software engineering. Paul Ammann and Jeff Offutt, *Introduction to Software Testing* (Cambridge: Cambridge University Press, 2016).
22. Riccardo Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* 51, no. 5 (2018): 93.
23. Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson, "Why Is My Robot Behaving Like That? Designing Transparency for Real Time Inspection of Autonomous Robots," in *AISB Workshop on Principles of Robotics* (Bath: Bath University Press, 2016).
24. Jan Leike et al., "AI Safety Gridworlds," *arXiv*, last revised November 28, 2017, arXiv:1711.09883.
25. Boutilier et al., "CP-nets."
26. Rossi, Venable, and Walsh, *A Short Introduction to Preferences*; Cornelio et al., "Reasoning with PCP-nets in a Multi-Agent Context"; Y. Chevaleyre et al., "Preference Handling in Combinatorial Domains: From AI to Social Choice," *AI Magazine* 29, no 4 (2008): 37–46; J. Goldsmith et al., "The Computational Complexity of Dominance and Consistency in CP-nets," *Journal of Artificial Intelligence Research* 33, no. 1 (2008): 403–32.
27. Yann Chevaleyre et al., "Learning Ordinal Preferences on Multiattribute Domains: The Case of CP-nets," in *Preference Learning*, ed. J. Fürnkranz and E. Hüllermeier, 273–96 (Springer, 2011); Allen, Thomas E., Muye Chen, Judy Goldsmith, Nicholas Mattei, Anna Popova, Michel Regenwetter, Francesca Rossi, and Cchristopher Zwilling. "Beyond Theory and Data in Preference Modeling: Bringing Humans into the Loop." In *Proceedings of the 4th International Conference on Algorithmic Decision Theory*, Editors Toby Walsh, Lexington, Kentucky, USA 2015, Springer LNCS 9346: 3-18.
28. Vincent Conitzer, Jérôme Lang, and Lirong Xia, "Hypercubewise Preference Aggregation in Multi-Issue Domains," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence,* Barcelona, Catalonia, Spain, Editor Toby Walsh (2011), 158–63; N. Mattei et al., "Bribery in Voting over Combinatorial Domains Is Easy," in *Proceedings of the 11th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Valencia, Spain, Editors: Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, Michael Winikoff, IFAAMAS Press, 1407 – 1408, (2012).
29. Sen, *Choice, Ordering, and Morality*; Thomson, "The Trolley Problem"; J. F. Bonnefon, A. Shariff, and I. Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352, no. 6293 (2016): 1573–76.
30. D. Copp, *The Oxford Handbook of Ethical Theory* (Oxford University Press, 2005).

31. Rossi, "Moral Preferences"; Bert Musschenga and Anton van Harskamp, eds., *What Makes Us Moral? On the Capacities and Conditions for Being Moral* (Springer, 2013).
32. Greene et al., "Embedding Ethical Principles in Collective Decision Support Systems."
33. A. Loreggia et al., "Preferences and Ethical Principles in Decision Making," in *Proceedings of the 1st AAAI/ACM Conference on AI, Ethics, and Society,* New Orleans, LA, USA, Editors: Jason Furman, Gary E. Marchant, Huw Price, Francesca Rossi, ACM, 222. (2018); A. Loreggia et al., "Value Alignment via Tractable Preference Distance," in *Artificial Intelligence Safety and Security*, ed. R. V. Yampolskiy 249 - 262 (CRC Press, Boca Raton, FL, USA 2018).
34. A. Loreggia et al., "On the Distance between CP-nets," in *Proceedings of the 17th International Joint Conference on Autonomous Agents and Multi-Agent Systems,* Stockholm, Sweden, Editors: Elisabeth André, Sven Koenig, Mehdi Dastani, Gita Sukthankar, International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM 955 – 963. (2018); A. Loreggia et al., "CPDist: Deep Siamese Networks for Learning Distances between Structured Preferences," *arXiv*, last revised June 20, 2019, arXiv:1809.08350.
35. Loreggia et al., "Preferences and Ethical Principles in Decision Making"; Loreggia et al., "Value Alignment via Tractable Preference Distance."
36. Charles Spearman, "The Proof and Measurement of Association between Two Things," *American Journal of Psychology* 15, no. 1 (1904): 72–101.
37. A. Loreggia, F. Rossi, and K. B. Venable, "Modelling Ethical Theories Compactly," in *Proceedings of the AAAI Workshop: AI, Ethics, and Society* (2017).
38. Loreggia et al., "On the Distance between CP-nets."
39. Loreggia et al., "Value Alignment via Tractable Preference Distance."
40. Dylan Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning," in *Proceedings of the 29th International Conference on Neural and Information Processing,* Barcelona, Spain, Editors: Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, Roman Garnett, Neural Information Processing Systems Foundation Inc., 3909 – 3917. (2016).
41. Loreggia et al., "On the Distance between CP-nets."
42. Loreggia et al., "Value Alignment via Tractable Preference Distance."
43. Greene et al., "Embedding Ethical Principles in Collective Decision Support Systems."
44. Goldsmith and Junker, "Preference Handling for Artificial Intelligence"; Chevaleyre et al., "Preference Handling in Combinatorial Domains."
45. Colin Allen, Gary Varner, and Jason Zinser, "Prolegomena to Any Future Artificial Moral Agent," *Journal of Experimental and Theoretical Artificial Intelligence* 12, no. 3 (2000): 251–61; Musschenga and van Harskamp, *What Makes Us Moral?*
46. Bonnefon, Shariff, and Rahwan, "The Social Dilemma of Autonomous Vehicles."
47. For example, Allen, Varner, and Zinser, "Prolegomena to Any Future Artificial Moral Agent."
48. Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (Cambridge: Cambridge University Press, 2011).
49. Loreggia, Rossi, and Venable, "Modelling Ethical Theories Compactly"; Rossi, "Moral Preferences."

50. Francesca Rossi, Peter van Beek, and Toby Walsh, eds., *Handbook of Constraint Programming* (Elsevier, 2006).
51. Pedro Meseguer, Francesca Rossi, and Thomas Schiex, "Soft Constraints," in *Handbook of Constraint Programming*, ed. Francesca Rossi, Peter van Beek, and Toby Walsh (Elsevier, 2006).
52. Ibid.
53. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W. H. Freeman, 1979).
54. Rossi, van Beek, and Walsh, *Handbook of Constraint Programming*.
55. Rina Dechter, *Constraint Processing* (Morgan Kaufmann, 2003).
56. Loreggia et al., "On the Distance between CP-nets"; Loreggia et al., "CPDist."
57. Boutilier et al., "CP-nets."
58. Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA: MIT Press, 2009).
59. Notice that "CP-nets" is written with capital *CP*, while "cp-statements" is written with small "cp," since they mean different things and since the literature introducing such notions uses this notation.
60. Boutilier et al., "CP-nets."
61. Ibid.
62. Goldsmith et al., "The Computational Complexity of Dominance and Consistency in CP-nets"; T. E. Allen et al., "Uniform Random Generation and Dominance Testing for CP-nets," *Journal of Artificial Intelligence Research* 59 (2017): 771–813.
63. Copp, *The Oxford Handbook of Ethical Theory*.
64. Deontologists sometimes believe in absolute constraints, that is, constraints that require a person to act a certain way no matter the consequences. But other deontologists do not. What separates deontologists from utilitarians is that deontologists believe in at least some constraints that can prevent us from pursuing the actions with the best outcomes.
65. William MacAskill, "Normative Uncertainty" (PhD thesis, University of Oxford, 2014).
66. Thomson, "The Trolley Problem"; Scalable Cooperation at MIT Media Lab, "Moral Machine," 2014.
67. Leike et al., "AI Safety Gridworlds."
68. Copp, *The Oxford Handbook of Ethical Theory*.
69. Bonnefon, Shariff, and Rahwan, "The Social Dilemma of Autonomous Vehicles."
70. Scalable Cooperation at MIT Media Lab, "Moral Machine." http://moralmachine.mit.edu/
71. Loreggia et al., "Value Alignment via Tractable Preference Distance"; Loreggia et al., "Preferences and Ethical Principles in Decision Making"; Loreggia et al., "On the Distance between CP-nets"; Loreggia et al., "CPDist."
72. C. Cornelio et al., "Updates and Uncertainty in CP-nets," in *Proceedings of the 26th Australasian Joint Conference on Artificial Intelligence,* Dunedin, New Zealand, Editors Stephen Cranefield and Abhaya C. Nayak, Springer LNCS 8272, 301 – 312. (2013); Cornelio et al., "Reasoning with PCP-nets in a Multi-Agent Context"; D.

Bigot et al., "Probabilistic Conditional Preference Networks," in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence Bellevue, WA, Editors Ann Nicholson and Padhraic Smyth, AUAI Press, 1-8,* (2013).

73. Francesca Rossi and Nicholas Mattei, "Building Ethically Bounded AI," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, AAAI Press,* 9785 – 9789. (2019).
74. Note that this preference is meant not to illustrate what a person would do but rather to illustrate the incomparability and other features of CP-nets.
75. Copp, *The Oxford Handbook of Ethical Theory*; Scalable Cooperation at MIT Media Lab, "Moral Machine."
76. M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika* 30, nos. 1–2 (1938): 81–93.
77. Ronald Fagin et al., "Comparing Partial Rankings," *SIAM Journal on Discrete Mathematics* 20, no. 3 (March 2006): 628–48.
78. Loreggia et al., "On the Distance between CP-nets."
79. Ibid.
80. Ibid.
81. Loreggia et al., "Value Alignment via Tractable Preference Distance."
82. Note that we could constrain this to be exactly if we wish.
83. Loreggia et al., "Value Alignment via Tractable Preference Distance"; Loreggia et al., "On the Distance between CP-nets."
84. One could argue that we should only judge the action, and not the overall preferences of the person. However, we chose to judge the whole person in this work, that is, their entire preference structure, because often in legal decisions we judge not only the action but also the intent and character of the person doing the action.
85. Greene, "The Cognitive Neuroscience of Moral Judgment and Decision Making."
86. Sen, *Choice, Ordering, and Morality*.
87. Chevaleyre et al., "Learning Ordinal Preferences on Multiattribute Domains."
88. Francesca Rossi and Alessandro Sperduti, "Learning Solution Preferences in Constraint Problems," *Journal of Experimental and Theoretical Artificial Intelligence* 10, no 1 (1998): 103–16; L. Khatib et al., "Solving and Learning a Tractable Class of Soft Temporal Constraints: Theoretical and Experimental Results," *AI Communications* 20, no. 3 (2007): 181–209.
89. Cornelio et al., "Updates and Uncertainty in CP-nets."
90. Ibid.
91. Loreggia et al., "On the Distance between CP-nets."

References

Allen, Colin, Gary Varner, and Jason Zinser. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12, no. 3 (2000): 251–61.

Allen, Thomas E., Muye Chen, Judy Goldsmith, Nicholas Mattei, Anna Popova, Michel Regenwetter, Francesca Rossi, and Cchristopher Zwilling. "Beyond Theory and Data in Preference Modeling: Bringing Humans into the Loop." In *Proceedings of the 4th International Conference on Algorithmic Decision Theory*, Editors Toby Walsh, Lexington, Kentucky, USA 2015, Springer LNCS 9346: 3-18.

Allen, Thomas E., Judy Goldsmith, Hayden Elizabeth Justice, Nicholas Mattei, and Kayla Raines. "Uniform Random Generation and Dominance Testing for CP-nets." *Journal of Artificial Intelligence Research* 59 (2017): 771–813.

Ammann, Paul, and Jeff Offutt. *Introduction to Software Testing*. Cambridge: Cambridge University Press, 2016.

Amor, Nahla Ben, Didier Dubois, Hela Gouider, and Henri Prade. "Graphical Models for Preference Representation: An Overview." In *Proceedings of the 10th International Scalable Uncertainty Management*, Nice, France, Editors Steven Schockaert, Pierre Senellart, Springer LNCS 9858 96–111. 2016.

Anderson Michael, and Susan Leigh Anderson. *Machine Ethics*. Cambridge: Cambridge University Press, 2011.

Scalable Cooperation at MIT Media Lab. "Moral Machine." 2014. http://moralmachine.mit.edu/

Balakrishnan, Avinash, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. "Incorporating Behavioral Constraints in Online AI Systems." In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, AAAI Press, 3 - 11*, 2019.

Balakrishnan, Avinash, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. "Using Contextual Bandits with Behavioral Constraints for Constrained Online Movie Recommendation." In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Editor Jérôme Lang, ijcai.org, 5802 – 5804, 2018.

Bigot, Damien, Helene Fargier, Jerome Mengin, and Bruno Zanuttini. "Probabilistic Conditional Preference Networks." In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, Bellevue, WA, Editors Ann Nicholson and Padhraic Smyth, AUAI Press, 1-8,* 2013.

Bistarelli, Stefano, Ugo Montanari, and Francesca. Rossi. "Semiring-Based Constraint Satisfaction and Optimization." *Journal of the ACM*, Volume 44, Number 2, 201 – 236, 1997.

Bonnefon, John-Francois, Azim Shariff, and Iyad Rahwan. "The Social Dilemma of Autonomous Vehicles." *Science* 352, no. 6293 (2016): 1573–76.

Boutilier, Craig, Ronen Brafman, Carmel Domshlak, Holger H. Hoos, and David. Poole. "CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements." *Journal of Artificial Intelligence Research* 21 (2004): 135–91.

Chevaleyre, Yann, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. "Preference Handling in Combinatorial Domains: From AI to Social Choice." *AI Magazine* 29, no 4 (2008): 37–46.

Chevaleyre, Yann, Frédéric Koriche, Jérôme Lang, Jérôme Mengin, and Bruno Zanuttini. "Learning Ordinal Preferences on Multiattribute Domains: The Case of CP-nets." In *Preference Learning*, edited by J. Fürnkranz and E. Hüllermeier, 273–96. Springer, Berlin 2011.

Conitzer, Vincent, Jérôme Lang, and Lirong Xia. "Hypercubewise Preference Aggregation in Multi-Issue Domains." In *Proceedings of the 22nd International Joint Conference on*

Artificial Intelligence, Barcelona, Catalonia, Spain, Editor Toby Walsh, IJCAI/AAAI Press, 158–63. 2011.

Copp, David *The Oxford Handbook of Ethical Theory*. Oxford University Press,Newyork. 2005.

Cornelio, Cristina, Judy Goldsmith, Nicholas Mattei, Francesca Rossi, and Kristen B. Venable. "Updates and Uncertainty in CP-nets." In *Proceedings of the 26th Australasian Joint Conference on Artificial Intelligence*, Dunedin, New Zealand, Editors Stephen Cranefield and Abhaya C. Nayak, Springer LNCS 8272, 301 – 312. 2013.

Cornelio, Cristina, Umberto Grandi, Judy Goldsmith, Nicholas Mattei, Francesca Rossi, and Kristen B. Venable. "Reasoning with PCP-nets in a Multi-Agent Context." In *Proceedings of the 14th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Istanbul, Turkey, Editors: Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind, ACM Press, 969 – 977. 2015.

Dechter, Rina. *Constraint Processing*. Morgan Kaufmann, 2003.

Domshlak, Carmel, Eyke Hüllermeier, Souhila Kaci, and Henri Prade. "Preferences in AI: An Overview." *Artificial Intelligence* 175, no. 7 (2011): 1037–52.

Fagin, Ronald, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. "Comparing Partial Rankings." *SIAM Journal on Discrete Mathematics* 20, no. 3 (March 2006): 628–48.

Fürnkranz Johannes, and Eyke Hüllermeier. *Preference Learning*. Springer, Berlin, 2010.

Garey Michael R., and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Goldsmith, Judy, and Ulrich. Junker. "Preference Handling for Artificial Intelligence." *AI Magazine* 29, no 4, 9 - 12 (2009).

Goldsmith, Judy, Jérôme Lang, Miroslaw Truszczyński, and Nic Wilson. "The Computational Complexity of Dominance and Consistency in CP-nets." *Journal of Artificial Intelligence Research* 33, no. 1 (2008): 403–32.

Gonzales, Christophe, and Patrice Perny. "GAI Networks for Utility Elicitation." In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning*, Whistler, Canada, Editors: Didier Dubois, Christopher A. Welty, Mary-Anne Williams, AAAI Press, 224 – 234. 2004.

Greene, Joshua. "The Cognitive Neuroscience of Moral Judgment and Decision Making." In *The Cognitive Neurosciences V,* edited by M. S. Cazzaniga. Cambridge, MA: MIT Press, 1 – 48. 2014.

Greene, Joshua, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. "Embedding Ethical Principles in Collective Decision Support Systems." In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, Editors: Dale Schuurmans and Michael P. Wellman, AAAI Press, 4147 – 4151. 2016.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51, no. 5 (2018): 93.

Hadfield-Menell, Dylan, Stuart J Russell, Pieter Abbeel, and Anca Dragan. "Cooperative Inverse Reinforcement Learning." In *Proceedings of the 29th International Conference on Neural and Information Processing*, Barcelona, Spain, Editors: Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, Roman Garnett, Neural Information Processing Systems Foundation Inc., 3909 – 3917. 2016.

Maurice. G. Kendall, M. G. "A New Measure of Rank Correlation." *Biometrika* 30, nos. 1–2 (1938): 81–93.

Khatib, Lina., Paul H. Morris, Robert A. Morris, Francesca Rossi, Alessandro Sperduti, Kristen B. Venable. "Solving and Learning a Tractable Class of Soft Temporal Constraints: Theoretical and Experimental Results." *AI Communications* 20, no. 3 (2007): 181–209.

Koller, Daphne, and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press, 2009.

Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI Safety Gridworlds." *arXiv*, last revised November 28, 2017. arXiv:1711.09883.

Loreggia, Andrea., Nicholas. Mattei, Francesca. Rossi, and Kristen B. Venable. "CPDist: Deep Siamese Networks for Learning Distances between Structured Preferences." *arXiv*, last revised June 20, 2019. arXiv:1809.08350.

Loreggia, Andrea., Nicholas. Mattei, Francesca. Rossi, and Kristen B. Venable. "On the Distance between CP-nets." In *Proceedings of the 17th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Stockholm, Sweden, Editors: Elisabeth André, Sven Koenig, Mehdi Dastani, Gita Sukthankar, International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM 955 – 963. 2018.

Loreggia, Andrea., Nicholas. Mattei, Francesca. Rossi, and Kristen B. Venable. "Preferences and Ethical Principles in Decision Making." In *Proceedings of the 1st AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans, LA, USA, Editors: Jason Furman, Gary E. Marchant, Huw Price, Francesca Rossi, ACM, 222. 2018.

Loreggia, Andrea., Nicholas. Mattei, Francesca. Rossi, and Kristen B. Venable. "Value Alignment via Tractable Preference Distance." In *Artificial Intelligence Safety and Security*, edited by R. V. Yampolskiy. CRC Press, Boca Raton, FL, USA 249 - 262 2018.

Loreggia, Andrea., Francesca. Rossi, and Kristen. B. Venable. "Modelling Ethical Theories Compactly." In *Proceedings of the AAAI Workshop: AI, Ethics, and Society*, San Francisco, CA, USA, AAAI Press, 2017.

MacAskill, William. "Normative Uncertainty." PhD thesis, University of Oxford, 2014.

Mattei, Nicholas, Maria Silvia Pini, Francesca Rossi, and Kristen B. Venable. "Bribery in Voting over Combinatorial Domains Is Easy." In *Proceedings of the 11th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Valencia, Spain, Editors: Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, Michael Winikoff, IFAAMAS Press, 1407 – 1408, 2012.

Mattei, Nicholas, and Toby Walsh. *PrefLib: A Library for Preferences*. Accessed February 21, 2020. http://www.preflib.org.

Mattei, Nicholas, and Toby. Walsh. "A PrefLib.Org Retrospective: Lessons Learned and New Directions." In *Trends in Computational Social Choice*, edited by U. Endriss, 289–309. AI Access Foundation, 2017.

Meseguer, Pedro, Francesca Rossi, and Thomas Schiex. "Soft Constraints." In *Handbook of Constraint Programming*, edited by Francesca Rossi, Peter van Beek, and Toby Walsh. Elsevier, 281 – 328, 2006.

Musschenga, Bert, and Anton van Harskamp, eds. *What Makes Us Moral? On the Capacities and Conditions for Being Moral*. Springer, Berlin, 2013.

Pigozzi, Gabriella, Alexis Tsoukiàs, and Paolo Viappiani. "Preferences in Artificial Intelligence." *Annals of Mathematics and Artificial Intelligence* 77 (2015): 361–401.

Ricci, Francesco, Lior Rokach, Bracha Shapira, and Paul B. Kantor, eds. *Recommender Systems Handbook*. Springer, 2011.

Rossi, Francesca "Moral Preferences." In *Proceedings of the 10th Workshop on Advances in Preference Handling at the International Joint Conference on Artificial Intelligence*, New York, New York, USA 2016.

Rossi, Francesca, and Nicholas Mattei. "Building Ethically Bounded AI." In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, AAAI Press,* 9785 – 9789. 2019.

Rossi, Francesca, and Alessandro Sperduti. "Learning Solution Preferences in Constraint Problems." *Journal of Experimental and Theoretical Artificial Intelligence* 10, no 1 (1998): 103–16.

Rossi, Francesca, Peter van Beek, and Toby Walsh, eds. *Handbook of Constraint Programming*. Elsevier, 2006.

Rossi, Francesca, Kristen B. Venable, and Toby Walsh. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Morgan and Claypool, 2011.

Sen, Amartya *Choice, Ordering, and Morality*. Blackwell, 1974.

Shoham, Yoav, and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge: Cambridge University Press, 2008.

Spearman, Charles. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15, no. 1 (1904): 72–101.

Stone, Peter. "Learning and Multiagent Reasoning for Autonomous Agents." In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyperbad, India, Editors: Manuela M. Veloso, IJCAI Press, 12–30, 2007.

Theodorou, Andreas, Robert H. Wortham, and Joanna J. Bryson. "Why Is My Robot Behaving Like That? Designing Transparency for Real Time Inspection of Autonomous Robots." In *AISB Workshop on Principles of Robotics*. Bath: Bath University Press, 2016.

Thomson, Judith Jarvis. "The Trolley Problem." *Yale Law Journal* 94, no. 6 (1985): 1395–415.

5

Computational Law, Symbolic Discourse, and the AI Constitution

Stephen Wolfram

5.1. Leibniz's Dream

Gottfried Leibniz, who died three hundred years ago, worked on many things. But a theme that recurred throughout his life was the goal of turning human law into an exercise in computation.[1] Of course, as we know, he didn't succeed. But three centuries later, I think we're finally ready to give it another serious try. And I think it's a really important thing to do—not just because it'll enable all sorts of new societal opportunities and structures, but because I think it's likely to be critical to the future of our civilization in its interaction with artificial intelligence.

Human law, almost by definition, dates from the very beginning of civilization, and undoubtedly it's the first system of rules that humans ever systematically defined. Presumably it was a model for the axiomatic structure of mathematics as defined by the likes of Euclid. And when science came along, "natural laws" (as their name suggests) were at first viewed as conceptually similar to human laws, except that they were supposed to define constraints for the universe (or God) rather than for humans.

Over the past few centuries we've had amazing success formalizing mathematics and exact science. And out of this there's a more general idea that's emerged: the idea of computation. In computation we're dealing with arbitrary systems of rules, not necessarily ones that correspond to mathematical concepts we know or features of the world we've identified. So now the question is: Can we use the ideas of computation, in very much the way Leibniz imagined, to formalize human law?

The basic issue is that human law talks about human activities, and (unlike, say, for the mechanics of particles) we don't have a general formalism for describing human activities. When it comes to talking about money, for example, we often can be precise, and as a result, it's pretty easy to write a very formal contract for paying a subscription or determining how an option on a publicly traded stock should work.

Stephen Wolfram, *Computational Law, Symbolic Discourse, and the AI Constitution* In: *Ethics of Artificial Intelligence.* Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0006

But what about all the things that typical legal contracts deal with? Well, clearly we have one way to write legal contracts: just use natural language (like English). It's often very stylized natural language because it's trying to be as precise as possible. But ultimately it's never going to be precise. Because at the lowest level it's always going to depend on the meanings of words, which for natural language are effectively defined just by the practice and experience of the users of the language.

5.2. A New Kind of Language

For a computer language, though, it's a different story. Because now the constructs in the language are absolutely precise: instead of having a vague, societally defined effect on human brains, they're defined to have a very specific effect on a computer. Of course, traditional computer languages don't directly talk about things relevant to human activities; they only directly talk about things like setting values for variables or calling abstractly defined functions.

But what I'm excited about is that we're starting to build a bridge between the precision of traditional computer languages and the ability to talk about real-world constructs. Actually it's something I've personally been working on for more than three decades now: our knowledge-based Wolfram Language.

The Wolfram Language is precise: everything in it is defined to the point where a computer can unambiguously work with it. But its unique feature among computer languages is that it's knowledge-based. It's not just a language to describe the low-level operations of a computer; instead, built right into the language is as much knowledge as possible about the real world. And this means that the language includes not just numbers like 2.7 and strings like "abc" but also constructs like the United States, the Consumer Price Index, and an elephant. And that's exactly what we need in order to start talking about the kinds of things that appear in legal contracts or human laws.

I should make it clear that the Wolfram Language as it exists today doesn't include everything that's needed. We've got a large and solid framework, and we're off to a good start. But there's more about the world that we have to encode to be able to capture the full range of human activities and human legal specifications.

The Wolfram Language has, for example, a definition of what a banana is, broken down by all kinds of details. So if one says "You should eat a banana," the language has a way to represent "a banana." But as of now, it doesn't have a meaningful way to represent "you," "should," or "eat."

Is it possible to represent things like this in a precise computer language? Absolutely! But it takes language design to set up how to do it. Language design is a difficult business—in fact it's probably the most intellectually demanding

thing I know, requiring a strange mixture of high abstraction together with deep knowledge and down-to-earth practical judgment. But I've been doing it now for nearly four decades, and I think I'm finally ready for the challenge of doing language design for everyday discourse.

So what's involved? Well, let's first talk about it in a simpler case: the case of mathematics. Consider the function Plus, which adds things like numbers together. When we use the English word "plus" it can have all sorts of meanings. One of those meanings is adding numbers together. But there are other meanings that are related, say, by various analogies ("product X plus," "the plus wire," "it's a real plus," etc.).

When we come to define Plus in the Wolfram Language we want to build on the everyday notion of "plus," but we want to make it precise. And we can do that by picking the specific meaning of "plus" that's about adding things like numbers together. Once we know that this is what Plus means, we immediately know all sorts of properties and can do explicit computations with it.

Now consider a concept like "magnesium." It's not as perfect and abstract a concept as Plus. But physics and chemistry give us a clear definition of the element magnesium, which we can then use in the Wolfram Language to have a well-defined "magnesium" entity.

It's very important that the Wolfram Language is a symbolic language because it means that the things in it don't immediately have to have "values"; they can just be symbolic constructs that stand for themselves. And so, for example, the entity "magnesium" is represented as a symbolic construct that doesn't itself "do" anything but can still appear in a computation, just like, for example, a number (like 9.45) can appear.

There are many kinds of constructs that the Wolfram Language supports, like "New York City," "last Christmas," and "geographically contained within." The point is that the design of the language has defined a precise meaning for them. "New York City," for example, is taken to mean the precise legal entity considered to be New York City, with geographical borders defined by law. Internal to the Wolfram Language is always a precise canonical representation for something like New York City (it's Entity["City", {"NewYork", "NewYork", "UnitedStates"}]). This internal representation is all that matters when it comes to computation. Yes, it's convenient to refer to New York City as "nyc," but in the Wolfram Language that natural language form is immediately converted to the precise internal form.

So what about "You should eat a banana"? Well, we've got to go through the same language design process for something like "eat" as for Plus (or "banana"). And the basic idea is that we've got to figure out a standard meaning for "eat." For example, it might be "ingestion of food by a person (or animal)." Now, there are plenty of other possible meanings for the English word "eat"—for example, ones that use analogies, as in "this function eats its arguments." But the idea—like for

Plus—is to ignore these and just to define a standard notion of "eat" that is precise and suitable for computation.

One gets a reasonable idea of what kinds of constructs one has to deal with just by thinking about parts of speech in English. There are nouns. Sometimes (as in "banana" or "elephant") there's a pretty precise definition of what these correspond to, and usually the Wolfram Language already knows about them. Sometimes it's a little vaguer but still concrete (as in "chair" or "window"), and sometimes it's abstract (like "happiness" or "justice"). But in each case one can imagine one or several entities that capture a definite meaning for the noun—just like the Wolfram Language already has entities for thousands of kinds of things.

Beyond nouns, there are verbs. There's typically a certain superstructure that exists around verbs. Grammatically there might be a subject for the verb, and an object, and so on. Verbs are similar to functions in the Wolfram Language: each one deals with certain arguments that, for example, correspond to its subject, object, and so on. Now of course in English (or any other natural language) there are all sorts of elaborate special cases and extra features that can be associated with verbs. But basically we don't care about these. Because we're really just trying to define symbolic constructs that represent certain concepts. We don't have to capture every detail of how a particular verb works; we're just using the English verb as a way to give us a kind of "cognitive hook" for the concept.

We can go through other parts of speech: adverbs that modify verbs; adjectives that modify nouns. These can sometimes be represented in the Wolfram Language by constructs like EntityInstance, and sometimes by options to functions. But the important point in all cases is that we're not trying to faithfully reproduce how the natural language works; we're just using the natural language as a guide to how concepts are set up.

Pronouns are interesting. They work a bit like variables in pure anonymous functions. In "You should eat a banana," the "you" is like a free variable that's going to be filled in with a particular person.

Parts of speech and grammatical structures suggest certain general features to capture in a symbolic representation of discourse. There are a bunch of others, though. For example, there are what amount to "calculi" that one needs to represent notions of time ("within the time interval," "starting later," etc.) or of space ("on top of," "contained within," etc.). We've already got many calculi like these in the Wolfram Language; the most straightforward are ones about numbers ("greater than," etc.) and sets ("member of"). Some calculi have long histories ("temporal logic," "set theory," etc.); others still have to be constructed.

Is there a global theory of what to do? Well, no more than there's a global theory of how the world works. There are concepts and constructs that are part of how our world works, and we need to capture these. No doubt there'll be new

things that come along in the future, and we'll want to capture those too. My experience from building Wolfram|Alpha is that the best thing to do is just to build each thing one needs, without starting off with any kind of global theory. After a while one may notice that one's built similar things several times, and one may go in and unify them.

One can get deep into the foundations of science and philosophy about this. Yes, there's a computational universe out there of all the possible rules by which systems can operate (and, yes, I've spent a good part of my life studying the basic science of this). And there's our physical universe that presumably operates according to certain rules from the computational universe. But from these rules can emerge all sorts of complex behavior; in fact the phenomenon of computational irreducibility implies that in a sense there's no limit to what can be built up.

But there's not going to be an overall way to talk about all this stuff. And if we're going to be dealing with any finite kind of discourse it's going to capture only certain features. Which features we choose to capture is going to be determined by what concepts have evolved in the history of our society. Usually these concepts will be mirrored in the words that exist in the languages we use.

At a foundational level, computational irreducibility implies that there'll always be new concepts that could be introduced. Back in antiquity, Aristotle introduced logic as a way to capture certain aspects of human discourse.[2] There are other frameworks that have been introduced in the history of philosophy, and more recently we have natural language processing and artificial intelligence research.[3] But computational irreducibility effectively implies that none of them can ever ultimately be complete. We must expect that as the concepts we consider relevant evolve, so too must the symbolic representation we have for discourse.

5.3. The Discourse of Workflow

Okay, so let's say we've got a symbolic representation for discourse. How's it actually going to be used? Well, there are some good clues from the way natural language works.

In standard discussions of natural language, it's common to talk about "interrogative statements" that ask a question, "declarative statements" that assert something, and "imperative statements" that say to do something. (Let's ignore "exclamatory statements," like expletives, for now.)

Interrogative statements are what we're dealing with all the time in Wolfram|Alpha: "What is the density of gold?"; "What is 3 + 7?"; "What was the latest reading from that sensor?" They're also common in notebooks used to interact with the Wolfram Language: there's an input (In[1]: = 2 + 2), and then there's a corresponding output (Out[1] = 4).

Declarative statements are all about filling in particular values for variables. In a very coarse way, one can set values (x = 7), as in typical procedural languages. But it's typically better to think about having environments in which one's asserting things. Maybe those environments are supposed to represent the real world, or some corner of it. Or maybe they're supposed to represent some fictional world, where, for example, dinosaurs didn't go extinct.

Imperative statements are about making things happen in the world: "Open the pod bay doors"; "Pay Bob 0.23 bitcoin."

In a sense, interrogative statements determine the state of the world; declarative statements assert things about the state of the world; and imperative statements change the state of the world.

In different situations, we can mean different things by "the world." We could be talking about abstract constructs, like integers or logic operations, that just are the way they are. We could be talking about natural laws or other features of our physical universe that we can't change. Or we could be talking about our local environment, where we can move around tables and chairs, choose to eat bananas, and so on. Or we could be talking about our mental states, or the internal state of something like a computer.

There are lots of things one can do if one has a general symbolic representation for discourse. But one of them—which is the subject of this chapter—is to express things like legal contracts. The beginning of a contract, with its various "whereas" clauses, recitals, definitions, and so on tends to be dense with declarative statements ("This is so"). Then the actual terms of the contract tend to end up with imperative statements ("This should happen"), perhaps depending on certain things determined by interrogative statements ("Did this happen?").

It's not hard to start seeing the structure of contracts as being much like programs. In simple cases, they just contain logical conditionals: "If X then Y." In other cases they're more closely modeled on math: "If this amount of X happens, that amount of Y should happen." Sometimes there's iteration: "Keep doing X until Y happens." Occasionally there's some recursion: "Keep applying X to every Y." And so on.

There are already some places where legal contracts are routinely represented by what amount to programs. The most obvious are financial contracts for things like bonds and options, which just amount to little programs that define payouts based on various formulas and conditionals.

There's a whole industry of using "rules engines" to encode certain kinds of regulations as "if-then" rules, usually mixed with formulas. In fact such things are almost universally used for tax and insurance computations. (They're also common in pricing engines and the like.)

Of course, it's no coincidence that one talks about "legal codes." The word "code," which comes from the Latin *codex*, originally referred to systematic

collections of legal rules. When programming came along a couple of millennia later, it used the word "code" because it basically saw itself as similarly setting up rules for how things should work, except now the things had to do with the operation of computers rather than the conduct of worldly affairs.

But now, with our knowledge-based computer language and the idea of a symbolic discourse language, we're trying to make it possible to talk about a broad range of worldly affairs in the same kind of way that we talk about computational processes. So we put all those legal codes and contracts into computational form.

5.4. Code versus Language

How should we think about symbolic discourse language compared to ordinary natural language? In a sense, the symbolic discourse language is a representation in which all the nuance and poetry have been crushed out of the natural language. The symbolic discourse language is precise, but it'll almost inevitably lose the nuance and poetry of the original natural language.

If someone says "2 + 2" to Wolfram|Alpha, it'll dutifully answer "4." But what if instead they say, "Hey, will you work out 2 + 2 for me?" Well, that sets up a different mood. But Wolfram|Alpha will take that input and convert it to exactly the same symbolic form as "2 + 2" and similarly just respond "4."

This is exactly the kind of thing that'll happen all the time with symbolic discourse language. If the goal is to answer precise questions—or, for that matter, to create a precise legal contract—it's exactly what one wants. One just needs the hard content that will actually have a consequence for what one's trying to do, and in this case one doesn't need the "extras" or "pleasantries."

Of course, what one chooses to capture depends on what one's trying to do. If one's trying to get psychological information, then the "mood" of a piece of natural language can be very important. Those "exclamatory statements" (like expletives) carry meaning one cares about. But one can still perfectly well imagine capturing things like that in a symbolic way, for example by having an "emotion track" in one's symbolic discourse language. (Very coarsely, this might be represented by sentiment or by position in an emotion space—or, for that matter, by a whole symbolic language derived, say, from emoji.)

In actual human communication through natural language, "meaning" is a slippery concept that inevitably depends on the context of the communication, the history of whoever is communicating, and so on. My notion of a symbolic discourse language isn't to try to magically capture the "true meaning" of a piece of natural language. Instead my goal is just to capture some meaning that one can then compute with.

For convenience, one might choose to start with natural language, and then try to translate it into the symbolic discourse language. But the point is for the symbolic discourse language to be the real representation; the natural language is just a guide for trying to generate it. In the end, the notion is that if one really wants to be sure one's accurate in what one's saying, one should say it directly in the symbolic discourse language, without ever using natural language.

Back in the 1600s, one of Leibniz's big concerns was to have a representation that was independent of which natural language people were using (French, German, Latin, etc.). One feature of a symbolic discourse language is that it has to operate "below" the level of specific natural languages.

There's a rough kind of universality among human languages, in that it seems to be possible to represent any human concept at least to some approximation in any language. But there are plenty of nuances that are extremely hard to translate—between different languages or the different cultures that surround them (or even the same language at different times in history). But in the symbolic discourse language, one's effectively "crushing out" these differences—and getting something that is precise, even though it typically won't correspond exactly to any particular human natural language.

A symbolic discourse language is about representing things in the world. Natural language is just one way to try to describe those things. But there are others. For example, one might give a picture. One could try to describe certain features of the picture in natural language ("a cat with a hat on its head"), or one could go straight from the picture to the symbolic discourse language.

In the example of a picture, it's very obvious that the symbolic discourse language isn't going to capture everything. Maybe it could capture something like "He is taking the diamond." But it's not going to specify the color of every pixel, and it's not going to describe all conceivable features of a scene at every level of detail.

In some sense, the symbolic discourse language is specifying a model of the system it's describing. And like any model, it's capturing some features and idealizing others away. But the importance of it is that it provides a solid foundation on which computations can be done, conclusions can be drawn, and actions can be taken.

5.5. Why Now?

I've been thinking about creating what amounts to a general symbolic discourse language for nearly forty years. But it's only recently, with the current state of the Wolfram Language, that I've had the framework to actually do it. And it's also

only recently that I've understood how to think about the problem in a sufficiently practical way.

Yes, it's nice in principle to have a symbolic way to represent things in the world. And in specific cases—like answering questions in Wolfram|Alpha—it's completely clear why it's worth doing this. But what's the point of dealing with more general discourse? For example, when do we really want to have a "general conversation" with a machine?

The Turing Test says that being able to do this is a sign of achieving general AI. But "general conversations" with machines—without any particular purpose in mind—so far usually seem in practice to devolve quickly into party tricks and Easter eggs. At least that's our experience looking at interactions people have with Wolfram|Alpha, and it also seems to be the experience with decades of chatbots and the like.

But the picture quickly changes if there's a purpose to the conversation, if you're actually trying to get the machine to do something or learn something from the machine. Still, in most of these cases, there's no real reason to have a general representation of things in the world; it's sufficient just to represent specific machine actions, particular customer service goals, or whatever. But if one wants to tackle the general problem of law and contracts, it's a different story. Because inevitably one's going to have to represent the full spectrum of human affairs and issues. And so now there's a definite goal to having a symbolic representation of the world: one needs it to be able to say what should happen and have machines understand it.

Sometimes it's useful to do that because one wants the machines just to be able to check whether what was supposed to happen actually did; sometimes one wants to actually have the machines automatically enforce or do things. But either way, one needs the machine to be able to represent general things in the world—and so one needs a symbolic discourse language to be able to do this.

5.6. Some History

In a sense, it's a very obvious idea to have something like a symbolic discourse language. Indeed it's an idea that's come up repeatedly across the course of centuries. But it's proved a very difficult idea to make work, and it has a history littered with (sometimes quite wacky) failures.

Things started well. Back in antiquity, logic as discussed by Aristotle provided a very restricted example of a symbolic discourse language.[4] When the formalism of mathematics began to emerge it provided another example of a restricted symbolic discourse language.

But what about more general concepts in the world? There'd been many efforts—between the Tetractys of the Pythagoreans and the I Ching of the Chinese—to assign symbols or numbers to a few important concepts.[5] But around 1300 Ramon Llull took it further, coming up with a whole combinatorial scheme for representing concepts—and then trying to implement this with circles of paper that could supposedly mechanically determine the validity of arguments, particularly religious ones.[6]

Four centuries later, Leibniz was an enthusiast of Llull's work, at first imagining that perhaps all concepts could be converted to numbers, and truth then determined by doing something like factoring into primes. Later Leibniz starting talking about a *characteristica universalis* (or, as Descartes called it, an "alphabet of human thoughts")[7]—essentially a universal symbolic language. But he never really tried to construct such a thing, instead chasing what one might consider "special cases"—including the one that led him to calculus.

With the decline of Latin as the universal natural language in the 1600s, particularly in areas like science and diplomacy, there had already been efforts to invent "philosophical languages" (as they were called) that would represent concepts in an abstract way, not tied to any specific natural language. The most advanced of these was by John Wilkins, who in 1668 produced a book cataloging over ten thousand concepts and representing them using strange-looking glyphs, with a rendering of the Lord's Prayer as an example.

In some ways these efforts evolved into the development of encyclopedias and later thesauruses, but as language-like systems they basically went nowhere. Two centuries later, though, as the concept of internationalization spread, there was a burst of interest in constructing new, country-independent languages, and out of this emerged Volapük and then Esperanto. These languages were really just artificial natural languages; they weren't an attempt to produce anything like a symbolic discourse language. I always enjoyed seeing signs in Esperanto at European airports and was disappointed when these finally disappeared in the 1980s. But, as it happens, right around that time, there was another wave of language construction. There were languages like Lojban, intended to be as unambiguous as possible, and ones like the interestingly minimal Toki Pona, intended to support the simple life, as well as the truly bizarre Ithkuil, intended to encompass the broadest range of linguistic and supposedly cognitive structures.

Along the way there were attempts to simplify languages like English by expressing everything in terms of one thousand or two thousand basic words (instead of the usual twenty thousand to thirty thousand), as in the "Simple English" version of Wikipedia or the xkcd Thing Explainer.

There were a few, more formal efforts. One was Hans Freudenthal's 1960 Lincos "language for cosmic intercourse" (i.e., communication with

extraterrestrials), which attempted to use the notation of mathematical logic to capture everyday concepts.[8] In the early days of the field of artificial intelligence, there were plenty of discussions of "knowledge representation," with approaches based variously on the grammar of natural language, the structure of predicate logic or the formalism of databases. Very few large-scale projects were attempted (Doug Lenat's Cyc being a notable counterexample).[9] When I came to develop Wolfram|Alpha I was disappointed at how little of relevance to our needs seemed to have emerged.

In a way I find it remarkable that something as fundamental as the construction of a symbolic discourse language should have had so little serious attention paid to it in the past. But at some level it's not so surprising. It's a difficult, large project, and it somehow lies in between established fields. It's not a linguistics project. Yes, it may ultimately illuminate how languages work, but that's not its main point. It's not a computer science project because it's really about content, not algorithms. And it's not a philosophy project because it's mostly about specific nitty-gritty and not much about general principles.

There've been a few academic efforts in the past half-century or so, discussing ideas like "semantic primes" and "natural semantic metalanguage."[10] Usually such efforts have tried to attach themselves to the field of linguistics, but their emphasis on abstract meaning rather than pure linguistic structure has put them at odds with prevailing trends, and none has turned into a large-scale project.

Outside of academia there's been a steady stream of proposals—sometimes promoted by wonderfully eccentric individuals—for systems to organize and name concepts in the world. It's not clear how far this pursuit has come since Ramon Llull—usually it's dealing only with pure ontology, and never with full meaning of the kind that can be conveyed in natural language.

I suppose one might hope that with all the recent advances in machine learning there'd be some magic way to automatically learn an abstract representation for meaning. One can take Wikipedia, for example, or a text corpus, and use dimension reduction to derive some effective "space of concepts." But, not too surprisingly, simple Euclidean space doesn't seem to be a very good model for the way concepts relate. (One can't even faithfully represent graph distances.) Even the problem of taking possible meanings for words—as a dictionary might list them—and breaking them into clusters in a space of concepts doesn't seem to be easy to do effectively.

Still, as I'll discuss later, I think there's a very interesting interplay between symbolic discourse language and machine learning. But for now my conclusion is that there's not going to be any alternative but to use human judgment to construct the core of any symbolic discourse language that's intended for humans to use.

5.7. Contracts into Code

Let's get back to contracts. Today there are hundreds of billions of them being signed every year around the world (and vastly more being implicitly entered into), though the number of "original" contracts that aren't just simple modifications is probably just in the millions (and is perhaps comparable to the number of original computer programs or apps being written).

Can these contracts be represented in precise symbolic form, as Leibniz hoped three hundred years ago? Well, if we can develop a decently complete symbolic discourse language, it should be possible. (Yes, every contract would have to be defined relative to some underlying set of "governing law" rules, etc., that are in some ways like the built-in functions of the symbolic discourse language.)

But what would it mean? Among other things, it would mean that contracts themselves would become computable things. A contract would be converted to a program in the symbolic discourse language. And one could do abstract operations just on this program. This means one can imagine formally determining—in effect through a kind of generalization of logic—whether, say, a given contract has some particular implication, could ever lead to some particular outcome, or is equivalent to some other contract.

Ultimately, though, there's a theoretical problem with this. Because questions like this can run into issues of formal undecidability, which means there's no guarantee that any systematic finite computation will answer them. The same problem arises in reasoning about typical software programs people write, and in practice it's a mixed bag, with some things being decidable, and others not.

Of course, even in the Wolfram Language as it is today, there are plenty of things (such as the very basic "Are these expressions equal?") that are ultimately in principle undecidable. And there are certainly questions one can ask that run right into such issues. But an awful lot of the kinds of questions that people naturally ask turn out to be answerable with modest amounts of computation. I wouldn't be surprised if this were true for questions about contracts too. (It's worth noting that human-formulated questions tend to run into undecidability much less often than questions picked, say, at random from the whole computational universe of possibilities.)

If one has contracts in computational form, there are other things one can expect to do, like to be able to automatically work out what the contracts imply for a large range of possible inputs. The 1980s revolution in quantitative finance started when it became clear one could automatically compute distributions of outcomes for simple options contracts. If one had lots (perhaps billions) of contracts in computational form, there'd be a lot more that could be done along these lines—and no doubt, for better or worse, whole new areas of financial engineering that could be developed.

5.8. Where Do the Inputs Come From?

Let's say one has a computational contract. What can one directly do with it? Well, it depends somewhat on the form of its inputs. One important possibility is that they're in a sense "born computational": they're immediately statements about a computational system ("How many accesses has this ID made today?"; "What is the ping time for this connection?"; "How much bitcoin got transferred?"). And in that case, it should be possible to immediately and unambiguously "evaluate" the contract—and find out if it's being satisfied.

This is something that's very useful for lots of purposes, both for humans interacting with machines and for machines interacting with machines. In fact there are plenty of cases where versions of it are already in use. Computer security provisions such as firewall rules is one example. Others are gradually emerging, such as automated SLAs (service-level agreements) and automated terms of service. (I'm certainly hoping our company, for example, will be able to make these a routine part of our business practices before too long.)

But it's certainly not true that every input for every contract is "born computational": plenty of inputs have to come from seeing what happens in the "outside" world ("Did the person actually go to place X?"; "Was the package maintained in a certain environment?"; "Did the information get leaked to social media?"; "Is the parrot dead?"). The first thing to say is that in modern times it's become vastly easier to automatically determine things about the world, not least because one can just make measurements with sensors. Check the GPS trace. Look at the car-counting sensor. And so on. The whole Internet of Things is out there to provide input about the real world for computational contracts.

Having said this, though, there's still an issue. Yes, with a GPS trace there's a definite answer (assuming the GPS is working properly) for whether someone or something went to a particular place. But let's say one's trying to determine something less obviously numerical. Let's say, for example, that one's trying to determine whether a piece of fruit should be considered "Fancy Grade" or not. Well, given some pictures of the piece of fruit an expert can pretty unambiguously tell. But how can we make this computational?

Here's a place where we can use modern machine learning. We can set up some neural net, say, in the Wolfram Language, and then show it lots of examples of fruit that's Fancy Grade and that's not. From my experience (and those of our customers!), most of the time we'll get a system that's really good at a task like grading fruit. It'll certainly be much faster than humans, and it'll probably be more reliable and more consistent too.

This gives us a whole new way to set up contracts about things in the world. Two parties can just agree that the contract should say "If the machine-learning system says X, then do Y." In a sense it's like any other kind of

computational contract: the machine-learning system is just a piece of code. But it's a little different because normally one expects that one can readily examine everything that a contract says; one can in effect read and understand the code. But with machine learning in the middle, there can no longer be any expectation of that.

Nobody specifically set up all those millions of numerical weights in the neural net; they were just determined by some approximate and somewhat random process from whatever training data were given. In principle we can measure everything about what's happening inside the neural net, but there's no reason to expect that we'll ever be able to get an understandable explanation—or prediction—of what the net will do in any particular case. Most likely it's an example of the phenomenon I call "computational irreducibility," which means there really isn't any way to see what will happen much more efficiently than just by running it.

What's the difference with asking a human expert, then, whose thought processes one can't understand? Well, in practice machine learning is much faster, so one can make much more use of "expert judgment." And one can set things up so they're repeatable, and one can, for example, systematically test for biases one thinks might be there, and so on.

Of course, one can always imagine cheating the machine learning. If it's repeatable, one could use machine learning itself to try to learn cases where it would fail. In the end it becomes rather like computer security, where holes are being found, patches are being applied, and so on. In some sense this is no different from the typical situation with contracts: one tries to cover all situations, then it becomes clear that something hasn't been correctly addressed, and one tries to write a new contract to address it, and so on.

The important bottom line is that with machine learning one can expect to get "judgment-oriented" input into contracts. I expect the typical pattern will be this: in the contract there'll be something stated in the symbolic discourse language (like "X will personally do Y"). And at the level of the symbolic discourse language there'll be a clear meaning to this, from which, for example, all sorts of implications can be drawn. But then there's the question of whether what the contract said is actually what happened in the real world. And, sure, there can be lots of sensor data that give information on this. But in the end a judgment call will have to be made: Did the person actually personally do this? Well—like for a remote exam-proctoring system—one can have a camera watching the person, one can record their pattern of keystrokes, and maybe even measure their EEG. But something's got to synthesize this data, make the judgment call about what happened, and turn this in effect into a symbolic statement. In practice I expect it will typically end up being a machine-learning system that does this.

5.9. Smart Contracts

Let's say we've got ways to set up computational contracts. How can we enforce them? Well, those that basically just involve computational processes can at some level enforce themselves. A particular piece of software can be built to issue licenses only in such-and-such a way. A cloud system can be built to make a download available only if it receives a certain amount of bitcoin. And so on.

How far do we trust what's going on? Maybe someone hacked the software or the cloud. How can we be sure nothing bad has happened? The basic answer is to use the fact that the world is a big place. As a (sometime) physicist, it makes me think of measurement in quantum mechanics. If we're just dealing with a little quantum effect, there's always interference that can happen. But when we do a real measurement, we're amplifying that little quantum effect to the point where so many things (atoms, etc.) are involved that it's unambiguous what happened—in much the same way as the Second Law of Thermodynamics makes it inconceivable that all the air molecules in a room will spontaneously line up on one side.

So it is with bitcoin, Ethereum, and so on. The idea is that some particular thing that happened ("X paid Y such-and-such" or whatever) is shared and recorded in so many places that there can't be any doubt about it. Yes, it's in principle possible that all the few thousand places that actually participate in something like bitcoin today could collude to give a fake result. But it's like with gas molecules in a room: the probability is inconceivably small. (As it happens, my Principle of Computational Equivalence suggests that there's more than an analogy with the gas molecules, and that actually the underlying principles at work are basically exactly the same. And, yes, there are lots of interesting technical details about the operation of distributed blockchain ledgers, distributed consensus protocols, etc., but I'm not going to get into them here.)

It's popular these days to talk about "smart contracts." When I've been talking about "computational contracts" I mean contracts that can be expressed computationally. But by "smart contracts" people usually mean contracts that can both be expressed computationally and execute automatically. Most often the idea is to set up a smart contract in a distributed computation environment like Ethereum, and then to have the code in the contract evaluate based on inputs from the computation environment.

Sometimes the input is intrinsic, like the passage of time (who could possibly tamper with the clock of the whole internet?) or physically generated random numbers. In cases like this, one has fairly pure smart contracts, say, for paying subscriptions or for running distributed lotteries.

More often there has to be some input from the outside—from something that happens in the world. Sometimes one just needs public information: the price

of a stock, the temperature at a weather station, or a seismic event like a nuclear explosion. Somehow the smart contract needs access to an "oracle" that can give it this information. And conveniently enough, there is one good such oracle available in the world: Wolfram|Alpha. And indeed Wolfram|Alpha is becoming widely used as an oracle for smart contracts. (Our general public terms of service say you currently just shouldn't rely on Wolfram|Alpha for anything you consider critical—though hopefully soon those terms of service will get more sophisticated, and computational.)

But what about nonpublic information from the outside world? The current thinking for smart contracts tends to be that one has to get humans in the loop to verify the information: in effect one has to have a jury (or a democracy) to decide whether something is true. But is that really the best one can do? I tend to suspect there's another path, that's like using machine learning to inject human-like judgment into things. Yes, one can use people, with all their inscrutable and hard-to-systematically-influence behavior. But what if one replaces those people in effect by AIs, or even a collection of today's machine-learning systems?

One can think of a machine-learning system as being a bit like a cryptosystem. To attack it and spoof its input one has to do something like inverting how it works. Given a single machine-learning system, there's a certain effort needed to achieve this. And if one has a whole collection of sufficiently independent systems, the effort goes up. It won't be enough just to change a few parameters in the system. But if one just goes out into the computational universe and picks systems at random, then I think one can expect to have the same kind of independence as by having different people. (To be fair, I don't yet quite know how to apply the mining of the computational universe that I've done for programs like cellular automata to the case of systems like neural nets.)

There's another point as well: if one has a sufficiently dense net of sensors in the world, then it becomes increasingly easy to be sure about what's happened. If there's just one motion sensor in a room, it might be easy to cover it. Maybe even if there are several sensors, it's still possible to avoid them, *Mission Impossible*–style. But if there are enough sensors, then by synthesizing information from them one can inevitably build up an understanding of what actually happened. In effect, one has a model of how the world works, and with enough sensors one can validate that the model is correct.

It's not surprising, but it always helps to have redundancy. More nodes to ensure the computation isn't tampered with. More machine-learning algorithms to make sure they aren't spoofed. More sensors to make sure they're not fooled. But in the end, there has to be something that says what should happen—what the contract is. And the contract has to be expressed in some language in which there are definite concepts. So somehow from the various redundant systems one has

in the world, one has to make a definite conclusion—one has to turn the world into something symbolic, on which the contract can operate.

5.10. Writing Computational Contracts

Let's say we have a good symbolic discourse language. How should contracts actually get written in it?

One approach is to take existing contracts written in English or any other natural language, and try to translate (or parse) them into the symbolic discourse language. What will happen is somewhat like what happens with Wolfram|Alpha today. The translator will not know exactly what the natural language was supposed to mean, and so it will give several possible alternatives. Maybe there was some meaning that the original writer of the natural-language contract had in mind. But maybe the poetry of that meaning can't be expressed in the symbolic discourse language: it requires something more definite. And a human is going to have to decide which alternative to pick.

Translating from natural-language contracts may be a good way to start, but I suspect it will quickly give way to writing contracts directly in the symbolic discourse language. Today lawyers have to learn to write legalese. In the future, they're going to have to learn to write what amounts to code: contracts expressed precisely in a symbolic discourse language.

One might think that writing everything as code rather than natural-language legalese would be a burden. But my guess is that it will actually be a great benefit. And it's not just because it will let contracts operate more easily. It's also that it will help lawyers think better about contracts. It's an old claim (the Sapir–Whorf hypothesis) that the language one uses affects the way one thinks. This is no doubt somewhat true for natural languages. But in my experience it's dramatically true for computer languages. Indeed I've been amazed over the years at how my thinking has changed as we've added more to the Wolfram Language. When I didn't have a way to express something, it didn't enter my thinking. But once I had a way to express it, I could think in terms of it.

So it will be, I believe, for legal thinking. When there's a precise symbolic discourse language, it'll become possible to think more clearly about all sorts of things.

Of course, in practice it'll help that there'll no doubt be all sorts of automated annotation: "If you add that clause, it'll imply X, Y and Z," for instance. It'll also help that it'll routinely be possible to take some contract and simulate its consequences for a range of inputs. Sometimes one will want statistical results ("Is this biased?"). Sometimes one will want to hunt for particular "bugs" that will be found only by trying lots of inputs.

Yes, one can read a contract in natural language, like one can read a math paper. But if one really wants to know its implications one needs it in computational form, so one can run it and see what it implies—and also so one can give it to a computer to implement.

5.11. The World with Computational Contracts

Back in ancient Babylon it was a pretty big deal when there started to be written laws like the Code of Hammurabi. Of course, with very few people able to read, there was a lot of clunkiness at first, like having people recite the laws in order from memory. Over the centuries things got more streamlined, and then about five hundred years ago, with the advent of widespread literacy, laws and contracts started to be able to get more complex (which, among other things, allowed them to be more nuanced and to cover more situations).

In recent decades the trend has accelerated, particularly now that it's so easy to copy and edit documents of any length. But things are still limited by the fact that humans are in the loop, authoring and interpreting the documents. Fifty years ago, pretty much the only way to define a procedure for anything was to write it down and have humans implement it. Then along came computers and programming. Very soon it started to be possible to define vastly more complex procedures—to be implemented not by humans, but instead by computers.

And so, I think, it will be with law. Once computational law becomes established, the complexity of what can be done will increase rapidly. Typically a contract defines some model of the world and specifies what should happen in different situations. Today the logical and algorithmic structure of models defined by contracts still tends to be fairly simple. But with computational contracts it'll be feasible for them to be much more complex, so that they can, for example, more faithfully capture how the world works.

Of course, that just makes defining what should happen even more complex—and before long it might feel a bit like constructing an operating system for a computer that tries to cover all the different situations the computer might find itself in.

In the end, though, one's going to have to say what one wants. One might be able to get a certain distance by just giving specific examples. But ultimately I think one's going to have to use a symbolic discourse language that can express a higher level of abstraction.

Sometimes one will be able to just write everything in the symbolic discourse language. But often, I suspect, one will use the symbolic discourse language to define what amount to goals, and then one will have to use machine-learning kinds of methods to fill in how to define a contract that actually achieves them.

As soon as there's computational irreducibility involved, it'll typically be impossible to know for sure that there are no bugs, or unintended consequences. Yes, one can do all kinds of automated tests, but in the end it's theoretically impossible to have any finite procedure that can guarantee to check all possibilities.

Today there are plenty of legal situations that are too complex to handle without expert lawyers. In a world where computational law is common, it won't just be convenient to have computers involved; it'll be necessary.

In a sense it's similar to what's already happened in many areas of engineering. Back when humans had to design everything themselves, they could typically understand the structures that were being built. But once computers are involved in design it becomes inevitable that they're needed in figuring out how things work too.

Today a fairly complex contract might involve a hundred pages of legalese. But once there's computational law—particularly contracts constructed automatically from goals—the lengths are likely to increase rapidly. At some level it won't matter, though, just as it doesn't really matter how long the code of a program one's using is. Because the contract will in effect just be run automatically by computer.

Leibniz saw computation as a simplifying element in the practice of law. And, yes, some things will become simpler and better defined. But a vast ocean of complexity will also open up.

5.12. What Does It Mean for AIs?

How should one tell an AI what to do? Well, you have to have some form of communication that both humans and AIs can understand—and that is rich enough to describe what one wants. As I've described elsewhere, this basically means that one has to have a knowledge-based computer language—which is precisely what the Wolfram Language is—and ultimately one needs a full symbolic discourse language.

But, okay, so one tells an AI to do something, like "Go get some cookies from the store." But what one says inevitably won't be complete. The AI has to operate within some model of the world and with some code of conduct. Maybe it can figure out how to steal the cookies, but it's not supposed to do that; presumably one wants it to follow the law or a certain code of conduct.

This is where computational law gets really important because it gives us a way to provide that code of conduct in a way that AIs can readily make use of.

In principle, we could have AIs ingest the complete corpus of laws and historical cases and so on, and try to learn from these examples. But as AIs become more and more important in our society, it's going to be necessary to define all

sorts of new laws, and many of these are likely to be "born computational," not least, I suspect, because they'll be too algorithmically complex to be usefully described in traditional natural language.

There's another problem too: we really don't want AIs to just follow the letter of the law (in whatever venue they happen to be); we want them to behave ethically too, whatever that may mean. Even if it's within the law, we probably don't want our AIs lying and cheating; we want them somehow to enhance our society along the lines of whatever ethical principles we follow.

One might think, why not just teach AIs ethics like we could teach them laws? In practice, it's not so simple. Because whereas laws have been somewhat decently codified, the same can't be said for ethics. Yes, there are philosophical and religious texts that talk about ethics, but they're a lot vaguer and less extensive than what exists for law.

Still, if our symbolic discourse language is sufficiently complete, it certainly should be able to describe ethics too. And in effect we should be able to set up a system of computational laws that defines a whole code of conduct for AIs.

But what should it say? One might have a few immediate ideas. Perhaps one could combine all the ethical systems of the world. Obviously hopeless. Perhaps one could have the AIs just watch what humans do and learn their system of ethics from that. Similarly hopeless. Perhaps one could try something more local, where the AIs switch their behavior based on geography, cultural context, and so on (think "protocol droid"). Perhaps useful in practice, but hardly a complete solution.

So what can one do? Well, perhaps there are a few principles one might agree on. For example, at least the way we think about things today, most of us don't want humans to go extinct. (Of course, maybe in the future, having mortal beings will be thought too disruptive, or whatever.) And actually, while most people think there are all sorts of things wrong with our current society and civilization, people usually don't want it to change too much, and they definitely don't want change forced upon them.

So what should we tell the AIs? It would be wonderful if we could just give the AIs some simple set of almost axiomatic principles that would make them always do what we want. Maybe they could be based on Asimov's Three Laws of Robotics. Maybe they could be something seemingly more modern, based on some kind of global optimization. But I don't think it's going to be that easy.

The world is a complicated place; if nothing else, that's basically guaranteed by the phenomenon of computational irreducibility. And it's pretty much inevitable that there's not going to be any finite procedure that'll force everything to come out the way one wants (whatever that may be).

Let me take a somewhat abstruse but well-defined example from mathematics. We think we know what integers are. But to really be able to answer all questions

about integers (including about infinite collections of them, etc.) we need to set up axioms that define how integers work. That's what Giuseppe Peano tried to do in the late 1800s.[11] For a while it looked good, but then in 1931 Kurt Gödel surprised the world with his Incompleteness Theorem, which implied, among other things, that try as one might, there was never going to be a finite set of axioms that would define the integers as we expect them to be and nothing else.[12]

In some sense, Peano's original axioms actually got quite close to defining just the integers we want. But Gödel showed that they also allow bizarre nonstandard integers, where, for example, the operation of addition isn't finitely computable.

That's abstract mathematics. What about the real world? One of the things that we've learned since Gödel's time is that the real world can be thought of in computational terms, pretty much just like the mathematical systems Gödel considered. In particular, one can expect the same phenomenon of computational irreducibility (which itself is closely related to Gödel's Theorem). The result is that whatever simple intuitive goal we may define, it's pretty much inevitable we'll have to build up what amounts to an arbitrarily complicated collection of rules to try to achieve it—and whatever we do, there'll always be at least some unintended consequences.

None of this should really come as much of a surprise. After all, if we look at actual legal systems as they've evolved over the past couple of thousand years, there always end up being a lot of laws. It's not like there's a single principle from which everything else can be derived; there inevitably end up being lots of different situations that have to be covered.

5.13. Principles of the World?

Is all this complexity just a consequence of the "mechanics" of how the world works? Imagine—as one expects—that AIs get more and more powerful. And that more and more of the systems of the world, from money supplies to border controls, are in effect put in the hands of AIs. In a sense, then, the AIs play a role a little bit like governments, providing an infrastructure for human activities.

So perhaps we need a constitution for the AIs, just like we set up constitutions for governments. But again the question comes up: What should the constitution have in it?

Let's say that the AIs could mold human society in pretty much any way. How would we want it molded? Well, that's an old question in political philosophy, debated since antiquity. At first an idea like utilitarianism might sound good: somehow maximize the well-being of as many people as possible. But imagine actually trying to do this with AIs that in effect control the world. Immediately one is thrust into concrete versions of questions that philosophers

and others have debated for centuries. Let's say one can sculpt the probability distribution for happiness among people in the world. Now we've got to get precise about whether it's the mean or the median or the mode or a quantile or, for that matter, the kurtosis of the distribution that we're trying to maximize.

No doubt one can come up with rhetoric that argues for some particular choice. But there just isn't an abstract right answer. We can have a symbolic discourse language that expresses any choice, but there's no mathematical derivation of the answer and there's no law of nature that forces a particular answer. I suppose there could be a "best answer given our biological nature." But as things advance, this won't be on solid ground either, as we increasingly manage to use technology to transcend the biology that evolution has delivered to us.

Still, we might argue, there's at least one constraint: we don't want a scheme where we'll go extinct and where nothing will in the end exist. Even this is going to be a complicated thing to discuss, however, because we need to say what the "we" here is supposed to be: just how "evolved" relative to the current human condition can things be, and not consider "us" to have gone extinct?

Even independent of this, there's another issue: given any particular setup, computational irreducibility can make it in a sense irreducibly difficult to find out its consequences. In particular, given any specific optimization criterion (or constitution), there may be no finite procedure that will determine whether it allows for infinite survival, or whether in effect it implies civilization will "halt" and go extinct.

So things are complicated. What can one actually do? For a little while there'll probably be the notion that AIs must ultimately have human owners, who must act according to certain principles, following the usual way human society operates. But realistically this won't last long.

Who would be responsible for a public-domain AI system that's spread across the internet? What happens when the bots it spawns start misbehaving on social media? (Yes, the notion that social media accounts are just for humans will soon look very "early twenty-first century.")

Of course, there's an important question of why AIs should "follow the rules" at all. After all, humans certainly don't always do that. It's worth remembering, though, that we humans are probably a particularly difficult case: we're the product of a multibillion-year process of natural selection, in which there's been a continual competitive struggle for survival. AIs are presumably coming into the world in very different circumstances, and without the same need for "brutish instincts." (I can't help thinking of AIs from different companies or countries being imbued by their creators with certain brutish instincts, but that's surely not a necessary feature of AI existence.)

In the end, though, the best hope for getting AIs to follow the rules is probably by more or less the same mechanism that seems to maintain human society

today: that following the rules is the way some kind of dynamic equilibrium is achieved. But even if we can get the AIs to follow the rules, we still have to define what the rules—the AI constitution—should be.

And, of course, this is a hard problem, with no right answer. But perhaps one approach is to see what's happened historically with humans. And one important and obvious thing is that there are different countries, with different laws and customs. So perhaps at the very least we have to expect that there'd be multiple AI constitutions, not just one.

Even looking at countries today, an obvious question is how many constitutions there should be. Is there some easy way to say that—with technology as it exists, for example—seven billion people should be expected to organize themselves into about two hundred countries?

It sounds a bit like asking how many planets the solar system should end up with. For a long time this was viewed as a "random fact of nature" (and widely used by philosophers as an example of something that, unlike 2 + 2 = 4, doesn't have to be that way). But particularly having seen so many exoplanet systems, it's become clear that our solar system actually pretty much has to have about the number of planets it does.

Maybe after we've seen the sociologies of enough video-game virtual worlds, we'll know something about how to derive the number of countries. But of course it's not at all clear that AI constitutions should be divided anything like countries.

The physicality of humans has the convenient consequence that at least at some level one can divide the world geographically. But AIs don't need to have that kind of spatial locality. One can imagine some other schemes, of course. Let's say one looks at the space of personalities and motivations and finds clusters in it. Perhaps one could start to say "Here's an AI constitution for that cluster," and so on. Maybe the constitutions could fork, perhaps almost arbitrarily (a "Git-like model of society"). I don't know how things like this would ultimately work, but they seem more plausible than what amounts to a single, consensus AI constitution for everywhere and everyone.

There are so many issues, though. Here's one: let's assume AIs are the dominant power in our world. But let's assume that they successfully follow some constitution or constitutions that we've defined for them. Well, that's nice—but does it mean nothing can ever change in the world? I mean, just think if we were still all operating according to laws that had been set up two hundred years ago; most of society has moved on since then and wants different laws (or at least different interpretations) to reflect its principles.

But what if precise laws for AIs were burned in around the year 2020, for all eternity? Well, one might say, real constitutions always have explicit clauses that allow for their own modification (in the US Constitution it's Article V). But looking at the actual constitutions of countries around the world isn't terribly

encouraging. Some just say basically that the constitution can be changed if some supreme leader (a person) says so. Many say that the constitution can be changed through some democratic process—in effect by some sequence of majority or similar votes. And some basically define a bureaucratic process for change so complex that one wonders if it's formally undecidable whether it would ever come to a conclusion.

At first, the democratic scheme seems like an obvious winner. But it's fundamentally based on the concept that people are somehow easy to count. (Of course, one can argue about which people, etc.) But what happens when personhood gets more complicated? When, for example, there are in effect uploaded human consciousnesses, deeply intertwined with AIs? One might say, there's always got to be some "indivisible person" involved. And yes, I can imagine little clumps of pineal gland cells that are maintained to define "a person," just like in the past they were thought to be the seat of the soul. But from the basic science I've done I think I can say for certain that none of this will ultimately work—because in the end the computational processes that define things just don't have this kind of indivisibility.

So what happens to democracy when there are no longer people to count? One can imagine all sorts of schemes, involving identifying the density of certain features in "people space." I suppose one can also imagine some kind of bizarre voting involving transfinite numbers of entities, in which perhaps the axiomatization of set theory has a key effect on the future of history.

It's an interesting question how to set up a constitution in which change is burned in. There's a very simple example in bitcoin, where the protocol just defines by fiat that the value of mined bitcoin goes down every year. Of course, that setup is in a sense based on a model of the world, and in particular on something like Moore's Law and the apparent short-term predictability of technological development. But following the same general idea, one might start thinking about a constitution that says "Change 1% of the symbolic code in this every year." But then one's back to having to decide "Which 1%?" Maybe it'd be based on usage or observations of the world or some machine-learning procedure. But whatever algorithm or meta-algorithm is involved, there's still at some point something that has to be defined once and for all.

Can one make a general theory of change? At first, this might seem hopeless. But in a sense exploring the computational universe of programs is like seeing a spectrum of all possible changes. And there's definitely some general science that can be done on such things. Maybe there's some setup—beyond just "Fork whenever there could be a change"—that would let one have a constitution that appropriately allows for change, as well as changing the way one allows for change, and so on.

5.14. Making It Happen

We've talked about some far-reaching and foundational issues, but what about the here and now? I think the exciting thing is that three hundred years after Leibniz died, we're finally in a position to do what he dreamed of: to create a general symbolic discourse language, and to apply it to build a framework for computational law.

With the Wolfram Language we have the foundational symbolic system—as well as a lot of knowledge of the world—to start from. There's still plenty to do, but I think there's now a definite path forward. It really helps that in addition to the abstract intellectual challenge of creating a symbolic discourse language, there's now also a definite target in mind: being able to set up practical systems for computational law.

It's not going to be easy, but I think the world is ready for it, and needs it. There are simple smart contracts already in things like bitcoin and Ethereum, but there's vastly more that can be done—and with a full symbolic discourse language the whole spectrum of activities covered by law becomes potentially accessible to structured computation. It's going to lead to all sorts of both practical and conceptual advances. And it's going to enable new legal, commercial, and societal structures—in which, among other things, computers will be drawn still further into the conduct of human affairs.

I think it's also going to be critical in defining the overall framework for AIs in the future. What ethics, and what principles, should they follow? How do we communicate these to them? For ourselves and for the AIs we need a way to formulate what we want. And for that we need a symbolic discourse language. Leibniz had the right idea, but three hundred years too early. Now in our time I'm hoping we're finally going to get to build for real what he only imagined. And in doing so we're going to take yet another big step forward in harnessing the power of the computational paradigm.

Notes

1. Stephen Wolfram, "Dropping In on Gottfried Leibniz," *Stephen Wolfram Writings,* May 14, 2013, writings.stephenwolfram.com/2013/05/dropping-in-on-gottfried-leibniz/.
2. Stephen Wolfram, *A New Kind of Science* (Wolfram Media, 2002), 860.
3. Ibid., 1103.
4. Ibid., 860.
5. Ibid., 1025.

6. Anthony Bonner, *The Art and Logic of Ramon Llull* (Leiden: Brill Academic, 2007), 290.
7. Richard A. Geiger and Brygida Rudzka-Ostyn, , eds., *Conceptualizations and Mental Processing in Language* (Walter de Gruyter, 1993), 25–26.
8. Wolfram, *A New Kind of Science*, 1189.
9. Cycorp, "Company Profile," www.cyc.com/company-profile/.
10. M. Lynne Murphy, *Lexical Meaning* (Cambridge: Cambridge University Press, 2010), 69–73
11. Wolfram, *A New Kind of Science*, 1189.
12. Ibid., 1158.

References

Bonner, Anthony. *The Art and Logic of Ramon Llull*. Leiden: Brill Academic, 2007.

Geiger, Richard A., and Brygida Rudzka-Ostyn, eds. *Conceptualizations and Mental Processing in Language*. Berlin: Walter de Gruyter, 1993.

Murphy, M. Lynne. *Lexical Meaning*. Cambridge: Cambridge University Press, 2010.

Wolfram, Stephen. "Dropping In on Gottfried Leibniz." *Stephen Wolfram Writings*, May 14, 2013. writings.stephenwolfram.com/2013/05/dropping-in-on-gottfried-leibniz/.

Wolfram, Stephen. *A New Kind of Science*. Champaign, IL: Wolfram Media, 2002.

Part II: The Near Future of Artificial Intelligence

6

Planning for Mass Unemployment

Precautionary Basic Income

Aaron James

Rapid automation leads to unemployment on a mass scale—not as a temporary shock, but as a new normal. Such was once the stuff of dystopian fiction. Now it could be our future, especially given amazing recent advances in machine learning and robotics. Not this year or this decade, necessarily. But down the road. After all, it's only dawn in the new machine age.[1]

Keynes warned in 1930 of a "new disease" that he called "technological unemployment," by which he meant "unemployment due to our discovery of means of economizing the use of labor outrunning the pace at which we can find new uses for labor."[2] Alas, his fears didn't materialize in much of the twentieth century. But by the close of the millennium there were signs that it's no iron law of economics that jobs destroyed will invariably bring new opportunities for human work. As Erik Brynjolfsson and Andrew McAffee explain, citing trends in computing capacity, digitization, and the power in recombining older technologies, "Employment grew alongside productivity up until the end of the twentieth century. [But then] job growth decoupled from productivity in the late 1990s."[3] Even as productivity rose, employment fell off. The long trend didn't hold. And so, in due course at least, artificial intelligence (AI) really might cause lasting structural unemployment on a mass scale.[4]

Imagine the scene: jobs are steadily automated, year after year. In the old days, for every job destroyed, a new one was eventually created, leaving total employment more or less unchanged. Now deep-learning machines, aided by clever entrepreneurs, race ahead and do the new tasks as well. The routine tasks, both manual and cognitive, have been mostly automated away. The best workers are those that work best with machines, in some cases besting both person and machine alike. A fair number still do what computers still can't (as of yet). But many workers—maybe most workers—are left to do the relative few manual or service jobs left. Wages have long stagnated or slipped. Joblessness used to be temporary for most; it wasn't fun, but the hand of technological fate was relatively arbitrary, spreading disruptions relatively evenly across the population, so that no one took the losses over and over again and nearly everyone shared the rising

Aaron James, *Planning for Mass Unemployment* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). © Oxford University Press. DOI: 10.1093/oso/9780190905040.003.0007

average standard of living. Now the lag between jobs grows ever longer. Many people—*most* people, even in their prime years—simply can't find tolerable work and stop looking. You wouldn't bet on mortgages, vacations, and college, rising wages over a lifetime, or better prospects for the children. It's enough just to try to get by. Reared in the Protestant Work Ethic, people are frustrated, distrustful, and angry, nursing contempt for the institutions that promised steady rewards for "hard work" and then failed to deliver decade after decade. With people driven ever farther apart by social media, and public discourse long since polluted, there's ever less scope for addressing common problems. Democracy, after its dramatic rise in the twentieth century, has become but a shell for corrupt authoritarianism, as hungry political entrepreneurs conspicuously feast on the unraveling.

It's an awful scenario. Relentless wage stagnation, the decline of trust, the hollowing out of democracy—that's all bad enough by itself, even without mass unemployment. Add millions of anxious souls out of work, and it's a terrible state of society—but now seemingly possible.

We certainly don't *know* work will be automated faster than reemployment any time soon. The economic historian Robert Gordon even calls the prospect "unlikely" (of which more later). Yet he has no doubt that a different slow-moving automation crisis is already upon us: the quality of the jobs there are has markedly declined.[5] Among the many factors (including trade, the decline of unions) that have brought stagnating wages since the mid-1970s, automation is an important part of the story.[6] He writes, "The problem created by the computer age is not mass unemployment but the gradual disappearance of good, steady, middle-level jobs that have been lost not just to robots and algorithms but to globalization and outsourcing to other countries, together with the concentration of job growth in routine manual jobs that offer relatively low wages."[7]

If nothing else, the idea that a mass unemployment crisis could eventually have awful social and political consequences, such as the rise of authoritarianism and hollowing out of democracy, shouldn't seem fanciful. The long trend of bad jobs and stagnant wages has arguably *already* had social and political consequences, in part for aiding the rise of authoritarianism and the hollowing out of democracy. Gordon thinks "this time is not different" as regards mass unemployment. But we're already in a *sort* of employment crisis that may warrant a precautionary basic income—precautionary, because things might get even worse.

And the different employment issues aren't in competition. The onset of mass technological unemployment would vastly worsen the bad jobs crisis we're in. We don't know we are proverbial frogs in boiling water, but caution seems to be in order, if only because of how bad things could become, if not next year or decade, then in the further future. By now it is not hard to imagine an upheaval

so radical as to frighten anyone conservative enough to value advanced peaceable civilization as we have come to know it (including those on the left who value what is by now a long history of progressive achievements). The risks to democracy alone are already grave enough. Recall that the historical rise of democracy is associated with the "long peace," a period of markedly reduced war and violence amid unprecedented prosperity. With the sunset of democracy, that trend might end.

That awful, entirely *possible* scenario might well belong in the same category of badness as a robot apocalypse or other catastrophic "existential" AI risks. Yet surely the risk of technological unemployment shouldn't have to be even *so* terrible for responsible people to see fit to take precautionary care. Even without mass unemployment, would not protracted wage stagnation (e.g., another *fifty* years of it) be bad enough? Is not the exacerbation of existing bad trends bad enough?

What to do now, if anything, calls for an ethical decision under uncertainty. Are the exciting benefits of a new machine age in the advanced countries—of increasing productivity, advances in medicine, greater convenience, and very cool gadgets—really worth the risk? What risks of disruption, distributed in what ways, should we find acceptable? How can we manage those risks? And shouldn't we somehow get a jump on the problem to ensure smooth passage into the new technological age?

As for what can be done, one apt and useful measure would be to establish a basic income grant—and do it *now*, or at the next available political opportunity, before a crisis of mass unemployment breaks out. A guaranteed minimum income would serve as a simple precautionary measure, not only to mitigate an eventual crisis but also to make it less likely. Or so I will explain. This chapter highlights the merits of taking this "precautionary approach" in comparison with less cautious options.

This argument is meant to be distinct from other familiar arguments for a basic income grant. There is, for example, the argument from sufficiency: that, for reasons of basic justice or basic humanity, everyone should be guaranteed a decent standard of living, because no one should live in poverty when this is avoidable. There is the egalitarian or prioritarian argument: that a basic income should go beyond mere "sufficiency" and advance a just distribution of welfare, freedom, opportunities, or social resources.[8] There is the feminist argument: that a basic income pays compensation for unpaid labor in reproduction and child care.[9] There is the republican argument: that a basic income reduces the risk of workers being "dominated" by employers, for having an improved bargaining position and a real chance of quitting or waiting longer for better work.[10] And there is the ecological or "green" argument, which sees basic income as fair return for the contribution of working less and doing something in leisure with

a lighter greenhouse gas footprint instead, such as playing games or sports or spending more time with the family.[11]

The argument from precaution in the face of technological risk is meant to be a distinct kind of argument with its own force. While it may draw on parallel considerations at various points, the argument is intended to answer a question about the justifiable imposition of risk for social benefit: What risks can we permissibly allow for the sake of technology's benefits?[12] The foregoing arguments for a basic income grant are not necessarily addressed to that question, and, for my purposes, all of them may fail on their own terms. My claim is that the "precautionary argument" is sufficient on its own. Establishing a basic income would be an important precautionary step to stave off and ameliorate a crisis of mass unemployment. That alone, I claim, is reason enough for us to set up a basic income grant today, with all deliberate speed, at least in the major advanced countries. (The risk is arguably less significant in the developing world, and perhaps in some rich countries as well. In that case my argument doesn't apply.)

6.1. Likely Enough?

Will self-reproducing psychopathic robots ultimately enslave humanity, or just end it by oblivious accident (e.g., in maximizing paper-clip production, as in Nick Bostrom's example)? Debates over these and other "existential" risks of AI tend to presume that we could well have to plan now for low-probability, catastrophic events.[13] The "AI safety" approach suggests precaution: technologists, firms, and governments should think hard, right away, *today*, about how to keep fast-moving developments under some kind of control, *before* the awful outcomes become at all likely.

It should seem odd that, by comparison, recent debates over technological unemployment have not had this precautionary focus. Skeptics about mass unemployment, in particular, have tended to debate the question of whether it is at all likely to happen in the future.[14] For example, Gordon explains in wonderful detail that, historically speaking, we do not find technological unemployment in the United States at least since the Civil War. He then argues that the big transformative productivity gains of the second Industrial Revolution—from low-hanging fruit in the internal combustion engine, the washing machine, electricity, sanitation—are behind us.[15] In the next most productive period, 1994–2004, the gains were much smaller and in any case but a "temporary upsurge that is unlikely to be repeated."[16] Any productivity gains from AI are unlikely to "match" the second Industrial Revolution, he says, at least within the next twenty-five or so years.[17] From this Gordon seems to conclude, without further argument, that we shouldn't worry about mass unemployment in particular.

Yet the real question is a different one: whether it's *credible enough* to justify taking policy action. And on that question Gordon is oddly silent.

It isn't that Gordon is a skeptic about futurology. He notes of the second Industrial Revolution that many of its unprecedented innovations were indeed predicted in 1863 by Jules Verne and in 1900 by none other than the *Ladies Home Journal*.[18] As for his own bets about AI, he simply never discusses the matter of basic credibility. He argues only that technological unemployment is unlikely in the relatively near future.

But, of course, the future may not be what it used to be, as Paul Krugman quips in his review of Gordon's book.[19] Technological unemployment has not been a problem *so far*. Could this time be different? It can seem at least a *credible* possibility, as Larry Summers seems to believe.[20] Is it credible enough that we should worry, watch, and start taking precautionary steps? Gordon does not say.

And why fixate on twenty-five years, as he does? Is there is still a *credible enough* chance of a new machine age and technological unemployment at some point in the next decades or half-century or longer? Even if a crisis is unlikely in the *near* future, it may be credible enough that *we shouldn't wait for it to become likely* over the longer run. Gordon doesn't sound worried, but he offers no reason to think history counsels *against* taking action.

To be sure, present-day statistics do not indicate a worrisome trend toward technological unemployment. After years of recovery after 2008, the United States is now near full employment. By one standard measure, output per worker per hour, productivity is not rising lately.[21] And we don't now find high levels of corporate investment in technology of the sort that would lead firms to eventually shed workers.

On the other hand, even historically, major productivity gains follow innovations with a long lag time.[22] How long might vary dramatically with circumstance. Workers are difficult to fire in good times, so one might expect cuts to appear mainly in recessions, perhaps the next one, or the next after.[23]

It would be a different matter if a crisis of mass unemployment were virtually impossible. A very *slight* possibility of a problem down the road would not justify precautionary action. We'd be wasting resources, being too cautious. For we can all agree that a significant, nontrivial expenditure on any possible problem is justified only when the potential bad outcome is at least credible enough, because our evidence and best guesses about it rise to some probability threshold. The question is: By what threshold? It could be fairly low, or there may be no general answer. It could be higher or lower depending on the badness of the outcome and our degree of aversion to risk.

When a bad outcome is *truly awful*, we should need less by way of credibility to sensibly pay something toward reducing its chance of happening. The possibility of an AI "singularity" run amok and the premature end to human

civilization is so utterly awful the mind freezes in the contemplation of it—perhaps so as to leave one sputtering about the point of anything if such an utterly foul demise comes to pass. It isn't just bad, horrible, and completely objectionable, but worse. Far worse. So one might conclude that, aside from whether or not it is likely, we really should do something about it, meaning now, before it's too late.

And is the chance of mass unemployment in the new machine age so incredible that we really shouldn't worry about it? It would seem rash to say so. Great revolutions become clear in the review mirror, in signs that appeared slowly, in the steady culmination of long trends, which coalesce in sudden changes. Much as with other historically unprecedented transformations, one wouldn't *expect* to be in a position to predict a second, even greater machine age revolution from the historical record. If some might have successfully read the tea leaves in the past, the farsighted can be distinguished from the cranks and dreamers only in retrospect. Again, a time frame of twenty-five years seems rather narrow. If a futurologist missed the target by only two or three or four decades, one would still rightly compliment his or her prescience.

There is indeed a long history of unfounded fears about technological unemployment. Perhaps one lesson of history is that we shouldn't listen to those fears. On the other hand, in light of the awesome history of human innovation, another lesson of history is that it is foolish to bet *against* technology. And one needn't believe an AI employment crisis is likely to refuse to rule it out as incredible or wholly unlikely; it can seem credible enough.

6.2. Precaution

Why buy a fire extinguisher for your home? You might know that your house's catching fire is very unlikely, statistically speaking. But you buy the extinguisher anyway as a precautionary measure, as you should, because (i) you really could wake in the night to your kitchen ablaze; (ii) this would be awful; and (iii) fire extinguishers are, given the stakes, cheap.

Once we're in a world of robot overlords, any chance to pull the plug will have long since passed. If only as a precautionary measure, then, we'd be wise to install an off switch, and make a general practice of it, from the start.[24] It isn't that today or next year is our last chance, necessarily, or that we won't know when to start risk management until it's too late. Whether or not it's likely, a robot apocalypse can be cause for precaution, much as with the fire extinguisher, when three jointly sufficient conditions for precaution are met: the adverse outcome need only be

1. Credible enough.
2. Terrible enough.
3. Open to precautionary steps that, given the high stakes, come cheap.

Neither an explosive AI singularity nor mass unemployment seem imminent as of the date I write this sentence in the year 2018. But if "safety" is on the table for the former, should we not consider it for the latter as well? Whether likely or not, the potential for an AI employment disruption—to people's lives, to the social fabric, and to democracy itself—is (i) credible enough, (ii) truly awful, and (iii) open to precautionary remedy on the cheap.

If or when AI does disrupt employment, one remedy is widely agreed upon: pay a basic income.[25] It'll be needed to prop up economic demand for the economy's basic functioning, not to mention to stave off mass hunger, homelessness, and violence in the streets. Arguably, we'd have to rewrite the social contract. With paid work in shrinking supply, we could professionalize part-time work, reduce the standard workweek, and encourage voluntary service and creative or sporting pursuits—all while defining ourselves less by our jobs and fashioning a new ethos of social contribution, meaning, and self-respect centered around things other than work.[26]

What's widely admitted is, however, only a rather modest point: that a basic income would be a necessary remedy *if* or *when* a crisis of mass technological unemployment breaks out. There is no consensus at all about its necessity *before* we see clear signs of it. Even those who support a basic income for other reasons (such as those listed earlier) might concede that, *at least for mass unemployment reasons*, a crisis should seem likely, or at least more likely than not. And since clear signs are not yet appearing in productivity statistics, a basic income remedy would then indeed have to wait. If we opted for it now, it would be for *other* reasons.

Yet at the dawn of a new machine age, should we not prepare ourselves? Why not do what will reduce the risks of disruption and ensure smooth passage into a supertechnological future? Along with the organized diminution of labor and other measures, a precautionary basic income, instituted with all deliberate speed, would go a long way to ensure smooth sailing on the changing technological seas.

Would this be too costly—costly enough to just wait and see? As I'll argue later, given the high stakes, the cost of a precautionary basic income would be eminently reasonable. Not quite as cheap as the wind, but a bargain nevertheless. For if you consider what we lose from taking precautionary adaptations—which is to say, the upside of sticking with business-as-usual—it's of limited importance. With so little to be gained, throwing caution to the wind is a bad gamble. It's like not buying a fire extinguisher when even pricey models are relatively cheap.

6.3. Precautionary Practice

The idea is that we can allow technological change to go forward, for the sake of its benefits, despite its risks. But instead of waiting for a crisis to loom or erupt, when we may not have the option of an orderly resolution, we take macroprudential steps now, very soon, in our next politically feasible chance, or at any rate well before signs of mass unemployment appear with any very high probability. We pay a basic income in order to prevent or at least mitigate a crisis, whenever, if ever, it might happen.

So it needn't be that we'll miss our last chance if we wait for a crisis to appear, though that might be true as well. The suggestion is rather that a basic income would reduce the chance of its very occurrence, and at any rate dampen its costs if it does occur. It would serve as both *prevention* and *mitigation*.

As mitigation, even with high unemployment, having a basic income would reduce its disruption to people's lives, the social fabric, trust, and the possibility of productive politics. Knowing one has money coming would cushion job losses and reduce anxieties, anger, and distrust. There'd be less blood on the water for demagogues staging a political attack.

As prevention, a basic income may reduce the amount of work that people want, and so make the employment level more robust. (The official rate tracks people looking for work.) For the greater security of an income guarantee, many might gradually become accustomed to working less, being happy to leave more of the human work available for others.

The suggestion is not unrealistic given current trends. Perhaps more and better opportunities are available for part-time work, as in present-day Switzerland, which allows workers to choose their hours without loss of professional stature or health benefits.[27] Or perhaps the workweek is limited to thirty-five hours, as in the case of the relatively "lazy" Germans.[28] In fact, a major German union just won wage hikes *and* the option of a twenty-eight-hour workweek over a two-year period (especially for child rearing), showing a clear preference for a better work-life balance.[29] Or maybe the workday is reduced to six hours for everyone, a mere thirty hours per week, as in Sweden's initiatives.[30] Even in the United States, the e-commerce giant Amazon, which isn't known for treating its workers well, has a pilot program that offers full benefits and a thirty-five-hour workweek at proportionally reduced pay.[31]

In principle, the workweek average could be lowered further as jobs become scarcer, eventually to twenty-five hours per week, and then twenty, if need be, until we reach some functional minimum. At each stage there are more opportunities for eager workers than would otherwise be available. So even with a decline in overall participation in the labor force, "full employment" would be easier to maintain. Even then a significant share of the population might still

work; perhaps most everyone works at least part time. So this isn't a "postwork" utopia—though there is that, if worse comes to worst.

And note that this does not require Luddite prohibitive taxes or regulations on innovation. Perhaps they won't ultimately curb a dangerous activity, or what is dangerous about it. But it is a major error to conclude, as Anti-Luddites tend to, that we are left with a choice between (i) doing nothing and (ii) trying to somehow compensate for market outcomes as cleanup, with whatever limited or symbolic measures are then available.

The idea of macro-prudential regulation is precisely to chart a safer middle course—to allow a dangerous activity to go forward, but then to address, ex ante, the level and distribution of its risks. This is the whole function of a precautionary basic income: it devises a social system in which each can expect net benefit from rising productivity, ex ante, because the risks of disruption have been sufficiently reduced.

It's the sort of public expectation that forms a social contract. Society can be said to have made a compact, in something like a promise to each worker. With expectations of steady prosperity formed and relied upon, workers ease in their strains of commitment, carrying on in good faith cooperation, living and working as society asks instead of sliding into alienated contempt and wanting to break something (while supporting politicians that therefore promise to break things).

To be sure, a precautionary basic income probably would not suffice on its own. In a world of growing unemployment, much more would be needed to adapt a culture that is now so thoroughly organized around work and striving. The possibility of taking *at least* that step raises the larger question: Why not retool work and the social contract now, in order to mitigate potential crises of employment, or even prevent them? Why wait for hell to break lose? I focus on basic income in this discussion, but here I note that similar issues arise for other ways of attuning the social contract to new conditions (such as climate change).[32]

That's my main argument for a precautionary approach, which I've stated in a preliminary way. I now turn to criticize some main alternatives.

6.4. Prudence

Consider a near worst-case scenario: joblessness spikes to as high as 75% over a decade because machines in abundant supply quickly and cheaply do anything a person can. As Brynjolfsson and McAfee suggest, a basic income would then surely be in nearly everyone's self-interest.[33] If people mainly have money though income from work, and too few have too little money in the larger population, there will be no expectation of profit, and the robot assets themselves will

soon be worth nothing. As the crisis worsens, the robot owner should therefore be happy to pony up taxes for a basic income to protect the value of his or her robot assets. And the unemployed certainly won't complain of being assured the basic means to live. Everyone benefits, so the matter of its ethical necessity may seem a side issue.

On the other hand, as a crisis builds, surely the calculating capitalist will want to know: If we robot owners are supposed to pay up, can't the *other* industries pay instead or first? Can I maybe drag my feet amid all the uncertainty about timing and scope of the problem? Shouldn't I try for the sweet spot in which the crisis eases and other industries or my industry competitors bear the cost of mitigation! And if each capitalist reasons thus, lobbying lawmakers and regulators accordingly, a basic income won't be set up.

The ethical question is therefore unavoidable. What we should say, instead, is that it is morally relevant that everyone benefits. In the imagined crisis, a basic income would be fully justifiable to everyone, including the capitalists. If even the investor and entrepreneurial class is better off paying the basic income taxes that keep them in business, then they can't reasonably object to it, not when it is in their own interest. (Many will still complain no doubt, but unreasonably.) And if the millions of workers at risk of personal ruin also have strong reasons to be assured of funds for food and a place to live, a basic income would be morally necessary.[34]

6.5. Laissez-faire

Even from an ethical perspective, there might be powerful reasons why the state should forgo all such schemes and simply let the chips fall where they may, for the sake of tech's long-run benefits. Then the question is: By what balance of prospects and risks?

The overall benefits are considerable, undoubtedly, especially as they accumulate in the long run. Yet what gains over which "long run" justify risks of great hardships to the whole life prospects of real people, perhaps to large swaths of society, over several generations? After all, in the "long run" we're all dead, as Keynes quipped. Should today's hard-working people be made to suffer for the sake of benefits to people yet to be born for thousands of years?

Here some will object tartly to Luddite prohibitions or taxes on innovation. As noted earlier, this is a red herring. Yes, the Luddites were wrong: *usually* when jobs are destroyed, new opportunities for work are created, preserving the overall employment level. And no, we wouldn't want to throw the growth baby out with the bathwater. The idea of a precautionary basic income is precisely that: to keep the growth baby, but without gratuitously throwing the livelihoods of millions of

people out as so much bathwater. And if that seems unworkable, consider that the United States, the rare country where a modest basic income is already established, could simply ramp up and broaden its existing Earned Income Tax Credit.

One traditional what-me-worry posture simply notes that technology has been a boon for humanity overall, and that rising standards of living amply justify occasional disruptions. But this is insufficient by itself. Nothing in the rising wealth of nations rules out macro-prudential planning or efforts to compensate "losers" in the name of Pareto efficiency or the general welfare. Indeed the case for free trade has always turned on some version of it.[35] Utilitarians such as Jeremy Bentham, J. S. Mill, John Harsanyi, and today's Peter Singer would indeed allow severe harm for the sake of long-term gains in the aggregate, at least in theory. But even they would also *require* short-run compensation through a perhaps generous basic income if that's best overall.[36]

A more sophisticated what-me-worry rationale for tech laissez-faire compares two world histories: our own, which has found great wealth and rising living standards for well over two centuries, and a world in which technological innovation was never left as free. Which would one choose? As Harsanyi would suggest, what you'd choose while assuming an equal chance of being anyone in any generation, in either history, while deciding from rational self-interest, is to live in the world as it is.

Yet this thought experiment uses an incomplete option set. Suppose, à la Harsanyi, that you could instead choose to live in a world of countries running basic incomes, again with an equal chance of being anyone in any generation, paid for with the ample gains from technological progress and trade. Then you'd definitely choose the basic income world over the world we're in. In an expected utility gamble, you'd go for the highest average prospect, and the beneficiaries of a basic income are a large share of the population. You wouldn't bet on being one of the relatively few capitalists who don't need the sure money.

6.6. Wait-and-See

A further approach admits that, yes, once an employment crisis breaks out, we would have to compensate those harmed with a basic income grant. But until such time as a new disruption becomes acute, or at least likely, perhaps more likely than not, we continue with business as usual.

This is to disregard the benefits of prevention, noted earlier. Paying a basic income ex ante can help *reduce the likelihood* that a crisis becomes acute, or even more likely than not. And if a crisis would be awful and can be made less likely

with reasonably cheap precautionary steps, it seems unwarranted to draw a sharp probability line at "more likely than not" or "acute" and so forth.

But prevention aside, why be at all confident in the feasibility of anything like "compensation" ex post? Once a crisis looms, it may be too late for an orderly resolution. Compare the infamous "Greenspan put," in which Fed Chairman Alan Greenspan chose not to "pull away the punch bowl" as a real estate bubble grew. Better to enjoy the froth, he thought, and clean up after the party is over. Alas, that gross miscalculation brought astronomical losses and the irreparable Great Recession mess. For loss of their jobs and homes and chances of getting on the same income escalator, millions were severely harmed. The scale of the damage was so large that any feasible program, at that point, would have been a token effort.[37]

In fact, in comparison to the scale of large bank bailouts, the measures to aid ordinary underwater homeowners were modest. Which again recalls unsavory politics: once a crisis of unemployment sets in, we should not feel confident that a window for arranging adequate compensation won't by that late date be closed.[38] For it is simply too easy, after a crisis, to rationalize away the need for compensation—because the crisis is but a "forty-year flood," a "black swan event," which was "beyond our control," or at any rate is now too large to do much of anything about. (And anyway, you know, "life is not fair, unfortunately.")

If we are planning ahead, on the other hand, there may well be a political opening to do what is anyway right. Accordingly, we'd be wise to take the next available political opportunity for lack of assurances that there'll be another. And if we pass on the next one, we'd be wise to take the next one after that, if there is one, and so on. Such matters of timing are highly uncertain, and we have no reason to expect fortuitous timing in political opportunities and real crises. So we should simply take the next political opportunity we have. Not because we know it will be our last but because we can't be sufficiently assured of having another.

One might doubt whether paying a basic income is the right precautionary measure to take right away. Why not simply commit now to better unemployment insurance, for instance, which pays out as people become unemployed? If people live on a flood plain, instead of paying them now, we can simply set up a flood insurance plan that compensates them when the deluge comes.

Here we should again be skeptical about whether such a plan would be faithfully followed ex post, after a crisis erupts. Will the promised payouts be seen as unworkable and unrealistic? Will the terrible outcomes be seen as unfortunate and beyond amelioration? If private insurance firms were involved, they may default, perhaps systematically, in the face of payment on such a large scale. They'd presumably require public backing, at the very least. But will all those promises

really be seen to remain in force? Even if only a fully public unemployment insurance scheme could credibly make payments on a vast scale, will even public promises be kept once a whole national economy is running at very high unemployment? Will they be vilified as bailouts which the industrious don't need and the rest don't deserve? Will things go down, in short, much as in the wake of 2008—this time with official promises renegotiated or "walked back," if not simply broken under a thin rationalizing pretext?

To be sure, any public promise or institution brings such political default risk, as one may call it. And robust unemployment insurance would certainly be better than nothing. If anything, however, the suggested difficulties highlight the importance of doing what we can to *prevent* a crisis from coming to pass in the first place. As emphasized earlier, paying a basic income ahead of time helps to *reduce the risks* that just such a crisis will break out. Expecting an insurance payout might bring some peace of mind, ahead of time. But more important, people who have enjoyed a basic income over a long period of time could be expected to adapt their work preferences and lifestyles as they go, instead of suffering a sudden shock of unemployment mitigated by unemployment checks. If there's steadily less demand for work, the total number of jobs can gradually shrink without producing a crisis.

Return to the people on the flood plain. If paid a basic income, then, much as with ordinary development, they're more likely to build sturdy homes and sewage systems that can withstand or at least adapt to heavy rains when they come. In general, what counts as a crisis is not simply a matter of an external shock (from bad weather or from AI), but our lack of preparedness for a major disruption. And steady money, especially over a long period of time, is precisely what might make people better prepared to work less and still find a contented way of life. (I return to this point momentarily.)

6.7. Ethical Discounts

But can we really afford a basic income grant, even just in the rich countries? The question of cost is indeed crucial. Though bad outcomes needn't be likely for us to reasonably take precautionary steps, if precaution is very expensive then perhaps we should wait and see until a crisis seems more likely, when the expenditure will be even better justified.

According to a very weak precautionary principle, inaction is not justified by the mere fact of our uncertainty. True enough, but here there *is* reason to think the cost and probability equation might change. We might start to see clear signs in productivity statistics. So perhaps we'd save in precautionary expense if we just wait. There is that potential upside of waiting to factor in.

A stronger precautionary principle, on the other hand, advises one to simply guard against highly consequential untoward outcomes, ignoring any potential upside of pressing our luck. In the present case, however, this thought *weakens* the case for precaution. By all means, let us consider the upside. The benefits of continued business as usual seem of limited value. What we lose in paying a precautionary basic income does have *some* value, but it should be sharply *discounted* from an ethical perspective. We'd have some extra money if we pass on the fire extinguisher, but it's not enough money to justify pushing our fire luck.

For starters, it's worth noting the possibility that a precautionary basic income might cost *nothing* in future growth. It might be part of an *optimal* growth path, for at least two conventional economic reasons. First, a basic income shifts resources toward people who have less, and hence they will have a higher marginal propensity to spend. So it amounts to a kind of economic stimulus. Second, speculation of a consequential sort tends to be done by the wealthier rather than the poorer, fueling boom-and-bust cycles that are extraordinarily costly over the longer run. So a shift in buying power away from the rich may reduce instability. These are abstract points of theory that can't be settled in the abstract. But in theory, at least, it may be that a reallocation of buying power would *optimize* growth over time. Then we can't afford *not* to set up a basic income.

Any such cost equation is of course highly controversial. So I will simply assume, *arguendo*, that a basic income would indeed come at a long-term net cost to economic growth. For example, perhaps we scale back work in anticipation of AI employment disruption, but more quickly than it turns out to be necessary to avoid a crisis. We then underproduce for decades on end. Despite increasing productivity, our growth path is not as high as it might otherwise have been, and we and our descendants don't become as rich as we otherwise would have on average.

But here we should ask: What is the *ethical* value of that opportunity forgone? How important is it, ethically speaking, for us to be even richer on average than we in the rich countries already are?

Potential growth certainly has great value in the developing world. High growth rates are part and parcel of the reduction of poverty. And for that, the advanced countries may need to retain some positive growth rate as well, if only for the benefits of international trade, which is key for the spread of technology.

That aside, what is the ethical value of advanced world growth in itself? How important is it for a country to become even richer on average, once it is rich already? If adopting a precautionary basic income now in the major rich countries means that we won't become as rich as we might have, how much is lost? The answer is: something, but not much.

This is for three main reasons. (1) The more we gain in GDP, the less and less it does for our happiness. (2) Work for GDP is expensive in time lost. (3) Further GDP gains have less value than comparable security benefits to the less well-off.

First, as utilitarians emphasize, gains in material wealth improve welfare less and less the richer we are. That's true on average, but it's especially true for the very rich. And the very rich now command a large share of any GDP gains.[39]

It's been said that money is at some point less about happiness than about "keeping score." Perhaps it is only human nature to in one way or another pursue ever more impressive symbols of one's worth in a status contest, if only to reassure oneself, just as Rousseau said. Yet as Rousseau would agree, even a money status game can equally be played on a lower playing field. The comparisons in relative status and worth can be made under *any* given overall standard of living, which might be set at a lower level according to ethical considerations. The standard level needn't rise without limit, and it can be tamped down with consumption taxes.[40] And if the very rich insist on competing for status over arbitrarily large and growing amounts of money, because only that seems at all interesting, then, well, the question is still whether we should roll the dice with technological fate just so they can keep their little party going.

Second, there is the opportunity cost of moneymaking in precious time lost. Further material goods become ever less meaningful for people given their limited time in life, because of the meaningful pursuits they must forgo in their pursuit. Not that *simply* having more time away from work must be especially great in itself; people get bored, or might have to endure a miserable state of unemployment, leaving them worse off than if they were working a nice job. But provided an adequate enough material standard, the value of time away from work, for rest, sport, socializing, voluntary service, community, and civic and political engagement indeed becomes ever more important. The opportunity cost of work as a mere means of moneymaking rises sharply in comparison.

That may be true even for those who love their work. The cost may be less important, but still something important they give up. For the many who must work but can't find work they love, and who have many good options outside of work, the cost is especially high. Today one is advised to finesse the trade-off, by doing something they love. But most people are not so fortunate, and not to be blamed for doing what they can in order to eat and pay the bills. It is an undesirable feature of a society that people are forced to find a way of monetizing the activities they love, simply to have enough time to pursue them vigorously. As Aristotle noted, the greatest joys in skillful activities are attained only with much practice, and they are felt with increasing force as one's competence or mastery increases. People may also find joy and meaning in maintaining a variety of skilled pursuits, which requires only more time, further raising the cost of mere moneymaking.

To be sure, we'll lose in GDP gains if we work less on average. But if those gains flow mainly to the very rich anyway, why work like crazy for *their* benefit? Why especially if even they are barely happier, if at all, for it? If *that's* what's so

important about growth, can everyone else really be asked to keep up long work hours, passing on time with the family at the lake or the beach?

Third, we should place more moral weight on the value of reducing risks of disruption to the relatively vulnerable. As we enter a new machine age, the *distribution* of risk among higher and lower skilled workers has intrinsic importance. A precautionary basic income would make the more vulnerable more secure. And that improvement is much more important, ethically speaking, than further riches to relatively well-off and secure people. That is, even if the rich *were* happier for making off with more than the lion's share of GDP gains (which they are not), the possibility of greater security and a better option for meaningful leisure for the relatively less well-off, including the middle class, has greater moral significance. The cost to the better off is relevant, but it counts for much less, ethically speaking. (I return to some doubts about this point later.)

6.8. A Great Bargain

A precautionary basic income is nevertheless expensive in budgetary terms. How will we pay for it?

One answer, which is admittedly controversial, is: monetary policy. Money is simply a transferrable promissory claim—an IOU—that is widely accepted in fulfilling a wide range of market obligations.[41] A sovereign that issues its own currency and requires that taxes be paid in its chosen unit of account can create new claims *by fiat*. The U.S. Central Bank now adds electronic zeroes to accounts that the major banks hold at the Central Bank, by so many keystrokes.[42] It can do the same by making electronic deposits (or by sending checks) to each citizen, paying a basic income from nothing, by simply deciding to do so.[43]

There is no "cost" to making these electronic credits, by keystroke. Or at least there's no cost beyond (i) any resulting inflation risks, in light of an appropriate target (which may need to be higher than current extremely low targets); (ii) any consequences for overall debt outlays (bearing in mind that no country that issues its own currency and borrowed in that denomination has defaulted, *ever*); and (iii) the opportunity cost for different valuable uses of monetary policy, which are then ruled out given whatever is required regarding (ii) and (iii).[44]

Since this use of monetary policy is highly controversial, I will assume, if only for the sake of argument, that a precautionary basic income must be paid for with taxes as well or instead. If income taxes are too controversial or anyway inadequate to the job, then financial, carbon, or consumption taxes and the like may be as or more desirable, subject to further considerations of efficiency and equity, whatever the overall tax level. (In fact VAT taxes are very successful in raising enormous sums.) Would the cost in extra taxes be worth it? Here my main claim

is simply that, since new taxes represent a cost to national production, including future growth, the cost should be steeply discounted for all of the ethical discount factors outlined earlier.

It's a further question how generous a basic income should be. The answer depends on what portion can be covered by monetary policy, and how far an increase in basic income would stimulate growth and pay for itself.[45] Assuming there is an upper bound on how high a basic income could be, my claim is that paying a basic income within this upper bound would be a bargain, given its considerable precautionary value. What's cheap or dear always depends on the payer's total budget and the value of what it otherwise might be used for. But even a large budgetary expenditure on a precautionary basic income can be *relatively cheap*, a real bargain. It can if it significantly reduces the chance of an especially bad outcome and the value of the expenditure is limited (as for the three reasons mentioned).

Should we worry that the robot invasion may never come? Maybe we take precaution and the crisis never happens. If things play out much as the techno-utopians dream, the cost of a precautionary basic income might seem wasted.

But if my argument is right so far, it isn't wasted any more than earthquake or auto insurance premiums are wasted if one is fortunate enough to never need to file an insurance claim. What one gained all along was *greater security*, the reduction of important risks, ex ante. The value of security remains important whatever happens ex post (though it is *especially* important if or when things do go south). And unlike setting up unemployment insurance, a precautionary basic income has the preventive function emphasized earlier.

Welcome to the AI casino. We have a gamble to take: we either continue with business as usual or adopt a precautionary basic income. Precaution may cost us something. But it offers smoother passage in the new machine age. So why take the extra risk, with so little to gain? Would it kill us to work less and spend more time with the family at the lake or the beach? Wouldn't most people be happier, as people themselves say in surveys? It's like the fire extinguisher or the earthquake or car insurance: you just wince and pay the premiums, gaining peace of mind in having protected yourself a bit better against the inevitable mishaps and disasters that come along with life.

6.9. Risk Equity

I suggested earlier that the case for precaution is strengthened once we weigh in the moral importance of risk distribution (the third ethical discount noted). The point is important, but also open to doubt even among those who might accept

the rest of the foregoing argument. Before concluding, I pay further attention to it here.

Automation creates differential risks. Those who do routinized work, whether in manual or cognitive labor, are at the greatest risk of unemployment. There's plenty of risk to go around, to be sure. Those who previously succeeded as a doctor or a lawyer now may take a backseat to those who work well with computers. Previously high-skilled industries may now require fewer and fewer employees. And some low-skilled, nonroutinized work, such as being a hair stylist, may prove robust. Yet lower skilled workers are *generally* more likely to be the ones who lose work, work for less, remain without work involuntarily, and see dim prospects. Everyone is at risk in case of swift mass technological employment, where the value even of robot assets collapses. But if we are all at some risk, capitalist and wage worker alike, the well-off will more easily ride out the crisis.

A precautionary basic income places a floor under everyone's prospects. Given these background differences, it also reduces the *relative* disparity in risk. It brings greater security to those now at greatest risk. For better distributing risk, it is not simply an optimal overall balance of risks and prospects, but equitable.

Or so I suggest. A utilitarian cost-benefit analysis, which might accept much of my argument, admits no reference to distribution as such. Each possible outcome of a policy choice (e.g., Jones loses work) is assessed one state at a time, without comparison to other possible states (Smith gets a raise).[46] The different outcomes, adjusted for their probability, are then aggregated (as a sum, average, or product) in different possible states of affairs, with no intrinsic, noninstrumental regard for how risks and prospects for different people compare.

It's true that a crisis of unemployment might matter for many reasons, personal, social, and political. So the question is: Does it also matter, per se, if Jones faces much sharper risks in *comparison* to Smith?

Not from behind Harsanyi's veil of ignorance. Suppose Jones is a spirited, hard-working, man of modest gifts. Behind Harsanyi's veil, he ignores these facts, supposing he has an equal chance of being anyone. As suggested earlier, if a large enough portion of the population is unemployed, his self-interested expected utility gamble would favor a world of basic income over laissez-faire, where the average prospect is higher. But, oddly, he'd equally be indifferent between life in this basic income world and a world of relatively high employment, where a smaller number of low-skilled workers scrape by with inconsistent bad jobs and repeated dislocations, without a guaranteed minimum. As long as they are few enough in number, the precariat can have very tough prospects. Jones is in fact one of those hardy people. But he won't have grounds for complaint about his having so much less than Smith—and nearly everyone else—if what's fair is settled behind Harsanyi's veil of ignorance.

But couldn't he reasonably ask for a basic income, at the very least? Perhaps, because fairness isn't well captured in Harsanyi's thought experiment. The issue goes to what rationality is, in the sense that's relevant to morality.

Harsanyi assumes coherence as defined in orthodox expected utility and applies it social choice.[47] It's the "sure thing" principle in particular, noted earlier, that says not to assess a possible outcome by comparison to other possible states. But this can seem to leave out morally relevant information. Consider Peter Diamond's well-known counterexample. A doctor has one donor kidney and two needy patients. He could simply give it to one of them, or he could flip a coin so that each has a chance. The sure thing principle treats these as equivalent, being indifferent between the coin toss and the doctor's simply picking one patient. (The expected value in each case is the same.) But, Diamond suggests, the options do seem morally different. The coin toss at least gives each a "fair shake."

It's hardly a marginal sort of case. As John Broome explains, it appears often in public choice: "For a given quantity of risk, is it better to have it more rather than less equally distributed? Radiation leaking from nuclear power stations will kill a number of people. Should nuclear policy be designed so that the risk of death is evenly distributed across the population, rather than concentrated on a smaller group of people? Suppose one hundred people will die; is it better to have ten million exposed to a .00001 chance of dying, than ten thousand exposed to a .01 chance? It is commonly believed that it is."[48]

To account for such cases, one can simply bake "fairness" into the consequence space.[49] This formally saves the principle but also raises a further question: Must what's morally at stake then be noncomparative in nature, such that it can be noncomparatively assessed and represented? Perhaps not, because the formal representation can be ad hoc, or simply describe moral conclusions reached by independent comparative reasoning. Or perhaps so, because any formal constraints should at least comport with relevant general features of moral evaluation, with no need for ad hoc fixes. In this case, however, the question is then what the general features of moral evaluation are. And quite aside from the present example, there is at least a plausible notion of how morality is *essentially* comparative.

In Scanlon's contractualism, for example, the interests of each kidney recipient must be assessed in comparative terms. Each of the would-be recipients has reason to live a bit longer, as well as an interest in being given a higher rather than a lower chance of receiving the kidney in the doctor's decision-making process. No comparison so far. But to figure out what each is *owed*, in all fairness, one compares the relative strength of each party's personal reasons in favor of one or another principle of doctor conduct, ranking them as complaints or objections by their overall force. If one of them is simply chosen at will, the other can object that she never had a chance. The coin toss, by comparison, respects their

symmetrical claims equally, giving each a 50-50 chance and so an equal, fair shake. If Diamond's example is suggestive, Scanlon's framework explains how it could reflect a structural feature of what we owe to each other. Morality, or a central part of it, is comparative through and through.[50]

It's a further question how the importance of risk distribution might be formally represented. As Matt Adler explains, it can be shoehorned into an aggregative cost-benefit framework as "distributive weights."[51] Adler also argues that a social welfare framework pinches less, allowing one to directly assess "risk equity" along with overall welfare.[52]sss

In that case, what's important now about a possible future of mass employment is that those who are relatively disadvantaged in society are subject to a greater share of the risk. It isn't just that the outcomes would be bad, should they happen. It's also that higher risks to the relatively disadvantaged are especially important. If the risks can be reduced ahead of time, it is especially hard to justify not taking precautionary action. A hard-working guy like Jones is getting a fairer shake.[53]

All this is to say that greater benefits to the less well-off matter to some extent. But by how much? John Rawls's difference principle imposes the demanding condition that we maximize expectations of the least well-off, at any cost to the better off that isn't necessary for this end.[54] Accordingly, behind Rawls's veil, parties decide under uncertainty, without assuming an equal chance of being anyone, and simply compare arrangements by their worst possible prospects and choose the one with the best worst ("maximin"). Contra Harsanyi's choosers, one would indeed choose a basic income world over a world of high employment, and more besides, since, for all one knows, one could well turn out to be Jones.

It's a famous and controversial argument. Here I simply note that "maximin" under uncertainty isn't necessary to express the prioritarian idea that greater benefits to the less well-off matter more—even behind a veil. As Laura Buchak has shown, even behind an equiprobability veil, a modestly risk-averse chooser will give priority to the less well-off. One need only give up the sure thing principle. A slightly heterodox expected utility framework allows one to give both welfare and equity their due.[55] How risk-averse the chooser should be dictates how much weight is given to greater benefits to the worse off. But this is a separate question. We can think it should be a *very significant* factor without saying exactly how significant it is in relative terms.

A nice consequence of this approach is that choice behind the veil becomes more like the morally significant risks we actually accept in real life.[56] Even a low-skilled worker like Jones might take a gamble on technological change, if presented with the right gamble. We all happily gamble on letting ambulances speed, taking a risk of being swiped and killed as we cross the street. We don't complain, even knowing that someone or other in a large population *will* take the hit. For we each stand to gain from expedited passage to a hospital ourselves,

and the chances to each of us of being killed prematurely is relatively slight. The license to speed is acceptable to all of us because we each stand to gain overall, ex ante, however the risks fall out.[57]

Likewise on the technological highway: it's one thing to be a low-skilled worker who is almost certain to suffer dearly so that others can enjoy a rising standing of living. That seems unfair. It's another thing—far more acceptable—if Jones's heightened risks of job loss is mitigated by assurance of a basic income. Then the practical certainty afforded by a decent floor might go a long way toward raising its ex ante acceptability. Jones at least then has hopeful prospects, given his own share of technology's general bounty. He might even feel better treated and able to relax a bit.

Economists such as Paul Samuelson have long supported free technological change. I suggest they assume roughly the foregoing normative picture of how risks are distributed.[58] But that picture isn't best captured by Harsanyi's utilitarianism, which isn't sensitive enough to the distribution of risk.[59]

6.10. Summary

I have not considered the many other forceful arguments for a basic income grant. If the reader still worries that the costs of precaution don't quite seem to justify action until mass unemployment looks likely, those other arguments might seal the case for establishing a basic income right away.

In any case I have argued that this position understates the good counsels of ethical macro-prudence. Farsightedness is not exactly in fashion lately. But those counsels are forceful and indeed quite sufficient on their own, even alongside other powerful (perhaps equally forceful) arguments. If in our anxious times an unemployment crisis is averted, in part for a few wise choices, eventually life for many really might be more like a day at the beach.[60]

Notes

1. Erik Brynjolfsson and Andrew McAffee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York: W. W. Norton, 2014). See also Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* (New York: Basic Books, 2015), and Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).
2. John Maynard Keynes, "Economic Possibilities for Our Grandchildren," in *Revisiting Keynes: Economic Possibilities of Our Grandchildren*, ed. Lorenzo Pecchi and Gustavo Piga (Cambridge, MA: MIT Press, 2008), 20-1 .

3. Brynjolfsson and McAffee, *The Second Machine Age*, 179–80.
4. See Daron Acemoglu and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets," NBER Working Paper No. 23285 (Cambridge, MA: National Bureau of Economic Research, 2017); Daron Acemoglu and Pascual Restrepo, "Robots and Jobs: Evidence from the US," Centre for Economic Policy Research, April 10, 2017, https://voxeu.org/article/robots-and-jobs-evidence-us; Melanie Arntz, Terry Gregory, and Ulrich Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis," OECD Social, Employment and Migration Working Papers (Paris: Organization of Economic Co-operation and Development, 2016).
5. Robert J. Gordon, *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War* (Princeton, NJ: Princeton University Press, 2016).
6. Ibid., 615.
7. Ibid., 604.
8. For the freedom variant, see Philippe Van Parijs, *Real Freedom for All: What (If Anything) Can Justify Capitalism?* (Oxford University Press, 1998)
9. Anne Alstott, "Good for Women," in *What's Wrong with a Free Lunch?*, eds. P. Van Parijs, J. Cohen and J. Rodgers (Boston: Beacon Press, 2001), 75-9.
10. Philip Pettit, "A Republican Right to Basic Income," *Basic Income Studies* 2, no. 2 (2007), 1-8.
11. Aaron James, *Surfing with Sartre: An Aquatic Inquiry into a Life of Meaning* (New York: Doubleday, 2017) , chs. 8–9, 226-257.
12. The question arises in many policy areas. I assume that it must be answered in light of the distinctive risks and prospects of technological change. For more general discussion of justifiable risk imposition and its distinctive argumentative demands, see Aaron James, "Contractualism's (Not So) Slippery Slope," *Legal Theory* 18, no. 3 (2012): 263–92; Aaron James, "The Distinctive Significance of Systemic Risk," *Ratio Juris* 30, no. 3 (2017), 239-58.
13. This claim is not always made explicitly. As Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 24 notes, he and others are betting on the rise of "superintelligence" within one hundred years, if not before. This is reported as a current best guess, however, and the overall argument does not require it. Likewise plausibility, not likelihood, is the focus in David Chalmers, "The Singularity," *Journal of Consciousness Studies* 17 (2010): 7–65, http://consc.net/papers/singularity.pdf.
14. Skeptics include Gordon, *The Rise and Fall of American Growth*; Tyler Cowen, *The Great Stagnation: How America Ate All the Low Hanging Fruit, Got Sick, and Will (Eventually) Feel Better* (New York: Penguin Press, 2011); and Lawrence Mishel and Josh Bivens, "The Zombie Robot Argument Lurches On," Economic Policy Institute, May 24, 2017, http://www.epi.org/files/pdf/126750.pdf.
15. Gordon, *The Rise and Fall of American Growth*, 575, figure 17-2.
16. Ibid., 602.
17. Ibid., 609. A similarly short time frame, of ten years, tends to be the focus in Mishel and Bivens, "The Zombie Robot Argument Lurches On."
18. Gordon, *The Rise and Fall of American Growth*, 590–91.

19. Paul Krugman, "Paul Krugman Reviews 'The Rise and Fall of American Grown' by Robert J. Gordon," *New York Times*, January 25, 2016, https://www.nytimes.com/2016/01/31/books/review/the-powers-that-were.html.
20. Larry Summers, "Economic Possibilities for Our Children," *NBER Reporter*, no. 4 (2013), 1-6.
21. Gordon, *The Rise and Fall of American Growth*, 604.
22. Ibid., 576.
23. Moreover, the standard ways of measuring an economy's output in GDP have well-known limitations. (Brynjolfsson and McAffee, *The Second Machine Age,* discuss several aspects of this problem on 115–25.) One of them is that they understate the value created in digital production. "Total factor productivity," which Gordon emphasizes, is famously difficult to measure. An improvement in a whole economy's ability to get more from the same resources—with innovations in techniques and technology, better organization, and other intangible factors—is not necessarily captured in unit-level measurement of a worker's output in a given hour, or the output per a unit of capital input, or a weighted average of these. As for what's now called the "Solow residual," Robert Solow quipped in 1987 that "We see the computer age everywhere, except in productivity statistics," Robert Solow, "We'd better watch out", *New York Times Book Review*, July 12 (1987), 36. He presumably meant to doubt the statistics, which hadn't by that time fastened on to emerging developments. So the sagging productivity statistics, though significant as data points, don't by themselves seem to justify any firm conclusion about whether apparent productivity stagnation is happening or likely to happen in the future.
24. Bostrom discusses various "control" measures and their limitations in *Superintelligence*, chs. 9–10.
25. Brynjolfsson and McAffee, *The Second Machine Age*, 232–41; Ford, *Rise of the Robots*, 257; Tegmark, *Life 3.0*, 127–29. Robert Solow is less explicit but suggests a need for the "democratization" of income in "Whose Grandchildren?," in *Revisiting Keynes: Economic Possibilities of Our Grandchildren*, ed. Lorenzo Pecchi and Gustavo Piga (Cambridge, MA: MIT Press, 2008), 87-93.
26. I describe the possibility and appeal of slow growth "leisure capitalism" in the advanced world in *Surfing with Sartre*, chs. 8–9, 226-257.
27. One person's experience with this is nicely described by Chantal Panozzo, "Living in Switzerland Ruined Me for America and Its Lousy Work Culture," *Vox*, February 1, 1016, https://www.vox.com/2015/7/21/8974435/switzerland-work-life-balance.
28. Annalyn Kurtz, "World's Shortest Work Weeks," *CNN Money*, July 10, 2013, https://money.cnn.com/gallery/news/economy/2013/07/10/worlds-shortest-work-weeks/5.html.
29. Alanna Petroff, "German Workers Win Right to 28-Hour Work Week," *CNN Money*, February 7, 2018, https://money.cnn.com/2018/02/07/news/economy/germany-28-hour-work-week/index.html.
30. Maddy Savage, "What Really Happened When the Swedes Tried Six-Hour Days?," *BBC News*, February 8, 2017, https://www.bbc.com/news/business-38843341.

31. Karen Turner, "Amazon Is Piloting Teams with a 30-Hour Workweek," *Washington Post*, August 26, 2016, https://www.washingtonpost.com/news/the-switch/wp/2016/08/26/amazon-is-piloting-teams-with-a-30-hour-work-week/?utm_term=.c96516b65c36.
32. In *Surfing with Sartre*, ch. 8, I give a similar argument for a basic income and reduction of the workweek from our obligation to mitigate the risks of climate change.
33. This is in effect the "Android Experiment" discussed in Brynjolfsson and McAffee, *The Second Machine Age*, 179.
34. The assumed view of morality here is T. M. Scanlon's account in *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998). In reasoning about right and wrong, we compare the reasons and claims of each party affected pair-wise, and then rank them (as "complaints" or "objections") by their force. The highest ranked objection to a principle of conduct counts as "reasonable" in comparison to other objections, settling what is required or permitted of us.
35. Albeit with a big fudge, depending on what view of "efficiency" is in question. While Pareto efficiency requires *actual compensation* and the establishment of social insurance institutions, Kaldor-Hicks efficiency requires only that the gains are large enough to *hypothetically* cover costs, even if compensation is never actually paid out. Economists have long ignored this crucial detail in order to make the case for free trade seem more decisive that it is. If the requisite social insurance schemes won't be established, then even free trade isn't efficient (in Pareto's sense). For discussion of the point and its significance, see Aaron James, *Fairness in Practice: A Social Contract for a Global Economy* (New York: Oxford University Press, 2012), ch. 2.
36. J. S. Mill urged as much in a similar connection, in his influential case for free trade: in dropping trade barriers, workers and other losers from imports should somehow be compensated.
37. For discussion, see the Bank of England's Andrew Haldane, "$100 Million Question," *Revista de Economía Institucional* 12, no. 22 (June 1, 2010), https://ssrn.com/abstract=1648936.
38. How will the robot owners lobby? Will an autocrat swept into power by a wave of discontent *feel* like offering a basic income grant? Maybe he's happy to call the unemployed worker "undeserving." Maybe the guy's a billionaire who rewards his elite loyalists and offers symbolic gestures to everyone else.
39. Matt Adler explains that standard cost-benefit analysis approaches do not take "diminishing marginal utility" into account. But it *should* be taken into account and can be by shifting to a social choice framework. Matthew Adler, *Well-being and Fair Distribution: Beyond Cost-Benefit Analysis* (Oxford: Oxford University Press, 2014).
40. Robert H. Frank, *Choosing the Right Pond: Human Behavior and the Quest for Status* (Oxford: Oxford University Press, 1985); Robert H. Frank, *The Darwin Economy: Liberty, Competition, and the Common Good* (Princeton, NJ: Princeton University Press, 2011).
41. Robert Hockett and Aaron James, *Money From Nothing: Or Why We Should Learn to Stop Worrying about Debt and Love the Federal Reserve* (New York: Melville House Books, 2020).

42. Here is Bernanke in an interview with interview for *60 Minutes*, in which he is asked how the recent open-market purchase of mortgage-based assets ("QE III") was paid for. Journalist Scott Pelley: "Is that tax money that the Fed is spending?" Bernanke: "It's not tax money. The banks have accounts with the Fed, much the same way that you have an account in a commercial bank. So, to lend to a bank, *we simply use the computer to mark up the size of the account* that they have with the Fed" (italics added).
43. In *Money From Nothing*, Robert Hockett and I propose new Central Bank tools and mandates, with equal independence from politics.
44. For discussion see L. Randall Wray, *Modern Money Theory: A Primer on Macroeconomics for Sovereign Monetary Systems*, 2nd ed. (New York: Palgrave Macmillan, 2015).
45. See Hockett and James, *Money From Nothing*.
46. This reflects the "sure thing principle," which Broome explains as follows: "The sure-thing requires outcomes to be assessed individually, one state at a time. But if there are interactions between states, they will not show up in a state-by-state assessment." John Broome, *Weighing Goods* (Oxford: Basil Blackwell, 1991), 110. I return to this momentarily.
47. John Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy* 61 (1953): 434–35. Similar assumptions about coherence (principle of irrelevant alternatives) inform Harsanyi's separate argument for utilitarianism, which does not rely on a veil of ignorance: John Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy* 63 (1955): 309–21.
48. Broome, *Weighing Goods*, 112.
49. This is Broome's solution (ibid., 113–14), which he defends on the pages following. Brian Skyrms (the decision theorist, game theorist, and philosopher) adopts the same solution (personal communication).
50. One could still keep sure thing as a principle of rational coherence but hold that it doesn't apply in the moral assessment of social choice. Or, because it's open to controversy in individual choice as well (see the Allais paradox), one can drop it entirely and modify the general expected utility framework, which might then apply to individual and social choice. On dropping the principle entirely, see Laura Buchak, *Risk and Rationality* (Oxford: Oxford University Press, 2013) and, for the application to social choice, Laura Buchak, "Taking Risks under the Veil of Ignorance," *Ethics* 127, no. 3 (2017), 610-44.
51. Which is to say "adjust costs and benefits with weighting factors that are inversely proportional to the well-being levels (as determined by income and also perhaps non-income attributes such as health) of the affected individuals." Matthew Adler, "Cost-Benefit Analysis and Distributional Weights: An Overview," Duke Environmental and Energy Economics Working Paper EE 13-04, August 2013, 1-29 https://ssrn.com/abstract=2467673.
52. Matthew Adler, "Risk Equity," *Harvard Environmental Law Review* 32, no. 1 (2008), 1-48.

53. Elsewhere I argue that being subject to high risk should *itself* be weighed in as intrinsically significant. I haven't suggested that here, but I take it to provide a still further reason for reducing risks to the less well off. See Aaron James, "The Distinctive Significance of Systemic Risk," *Ratio Juris* 30, no. 3 (2017): 239–58, https://ssrn.com/abstract=3016883.
54. This is the "lexical" version, which may not be Rawls's considered view. His case for the difference principle in John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971) arguably depends on contingencies such as "chain-connection" and "close-knittedness," which mean that classes rise and fall together.
55. See Buchak, "Taking Risks under the Veil of Ignorance," along with her *Risk and Rationality*. In John Rawls, *Justice as Fairness: A Restatement* (Cambridge, MA: Harvard University Press, 2001), Rawls accepts the framework of expected utility theory when it is understood to have "no substantive content" (99). He also says maximin isn't necessary except as a "useful heuristic device" under specified conditions (99).
56. This answers Harsanyi's criticism of Rawls's use of maximin as too different from everyday rationality. J. C. Harsanyi, "Can the Maximin Principle Serve as a Basis for Morality," *American Political Science Review* 69 (1975): 594–606.
57. The example is due to Frances Kamm. For development of the idea within ex ante contractualism, see James, "Contractualism's (Not So) Slippery Slope"; Johan Frick, "Contractualism and Social Risk," *Philosophy and Public Affairs* 43, no. 3 (2015): 175–223; Rahul Kumar, "Risking and Wronging," *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 27–51.
58. According to Samuleson's "heuristic theorem," "Most technical changes or policy choices directly help some people and hurt others. For some changes, it is possible for the winners to buy off the losers so that everyone could conceivably end up better off than in the prior status quo. Suppose that no such compensatory bribes or side payments are made, but assume that we are dealing with numerous inventions and policy decisions that are quasi-independent. Even if for each single change it is hard to know in advance who will be helped and who will be hurt, in the absence of known 'bias' in the whole sequence of changes, there is some vague presumption that a hazy version of the law of large numbers will obtain: so as the number of quasi-independent events becomes larger and larger, the chances improve that any random person will be on balance benefitted by a social compact that lets events take place that push out society's utility possibility frontier, even though any one of the events may push some people along the new frontier in a direction less favorable than the status quo."Paul Samuelson, "Bergsonian Welfare Functions," in *Economic Welfare and the Economics of Soviet Socialism*, ed. Steven Rosefielde (Cambridge: Cambridge University Press, 1981), 227. See also Edith Stokey and Richard Zeckhauser, *A Primer for Policy Analysis* (New York: W. W. Norton and Co., 1978), 283.
59. See Johann Frick's vaccine cases in "Contractualism and Social Risk." Harsanyi's veil does not distinguish between the "doomed" children in "Vaccine 3" and the merely "luckless" children in "Vaccine 2." An ex ante contractualism that uses a "natural veil

of ignorance" does (see 189–91). It is in this way sensitive to the difference between the widely shared risks in Vaccine 2 and the highly concentrated risks in Vaccine 3.

60. Thanks to my UCI seminar in winter 2016, cotaught with Brian Skyrms, and my UCI seminar in winter 2018 on technological and climate change. I have also benefited from conversations with Robert Hockett and Greg Shaffer.

References

Acemoglu, Daron, and Pascual Restrepo. "Robots and Jobs: Evidence from the US." Centre for Economic Policy Research, April 10, 2017. https://voxeu.org/article/robots-and-jobs-evidence-us.

Acemoglu, Daron, and Pascual Restrepo. "Robots and Jobs: Evidence from US Labor Markets." NBER Working Paper No. 23285. Cambridge, MA: National Bureau of Economic Research, 2017.

Adler, Matthew. "Cost-Benefit Analysis and Distributional Weights: An Overview." Duke Environmental and Energy Economics Working Paper EE 13-04, August 2013. https://ssrn.com/abstract=2467673.

Adler, Matthew. "Risk Equity." *Harvard Environmental Law Review* 32, no. 1 (2008).

Adler, Matthew. *Well-being and Fair Distribution: Beyond Cost-Benefit Analysis.* Oxford University Press, 2014.

Alstott, Anne. "Good for Women." In *What's Wrong with a Free Lunch?*, edited by edited by P. Van Parijs, J. Cohen and J. Rodgers (Boston: Beacon Press, 2001), 75-9.

Arntz, Melanie, Terry Gregory, and Ulrich Zierahn. "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis." OECD Social, Employment and Migration Working Papers. Paris: Organization of Economic Co-operation and Development, 2016.

Bostrom, Nick. *Superintelligence.* Oxford: Oxford University Press, 2014.

Broome, John. *Weighing Goods.* Oxford: Basil Blackwell, 1991.

Brynjolfsson, Erik, and Andrew McAffee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* New York: W. W. Norton, 2014.

Buchak, Laura. *Risk and Rationality.* Oxford: Oxford University Press, 2013.

Buchak, Laura. "Taking Risks under the Veil of Ignorance." *Ethics* 127, no. 3 (2017), 610-44.

Chalmers, David. "The Singularity." *Journal of Consciousness Studies* 17 (2010): 7–65, http://consc.net/papers/singularity.pdf.

Cowen, Tyler. *The Great Stagnation: How America Ate All the Low Hanging Fruit, Got Sick, and Will (Eventually) Feel Better.* New York: Penguin Press, 2011.

Ford, Martin. *Rise of the Robots: Technology and the Threat of a Jobless Future.* New York: Basic Books, 2015.

Frank, Robert H. *Choosing the Right Pond: Human Behavior and the Quest for Status.* Oxford: Oxford University Press, 1985.

Frank, Robert H. *The Darwin Economy: Liberty, Competition, and the Common Good.* Princeton, NJ: Princeton University Press, 2011.

Frick, Johan. "Contractualism and Social Risk." *Philosophy and Public Affairs* 43, no. 3 (2015): 175–223.

Gordon, Robert J. *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War.* Princeton, NJ: Princeton University Press, 2016.

Haldane, Andrew. "$100 Million Question." *Revista de Economía Institucional* 12, no. 22 (June 1, 2010), https://ssrn.com/abstract=1648936.

Harsanyi, J. C. "Can the Maximin Principle Serve as a Basis for Morality." *American Political Science Review* 69 (1975): 594–606.

Harsanyi, John. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (1953): 434–35.

Harsanyi, John. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63 (1955): 309–21.

Hockett, Robert and Aaron James. *Money From Nothing: Or Why We Should Learn to Stop Worrying about Debt and Love the Federal Reserve* (New York: Melville House Books, 2020).

James, Aaron. "Contractualism's (Not So) Slippery Slope." *Legal Theory* 18, no. 3 (2012): 263–92.

James, Aaron. "The Distinctive Significance of Systemic Risk." *Ratio Juris* 30, no. 3 (2017): 239–58, https://ssrn.com/abstract=3016883.

James, Aaron. *Fairness in Practice: A Social Contract for a Global Economy.* New York: Oxford University Press, 2012.

James, Aaron. *Surfing with Sartre: An Aquatic Inquiry into a Life of Meaning.* New York: Doubleday, 2017.

Keynes, John Maynard. "Economic Possibilities for Our Grandchildren." In *Revisiting Keynes: Economic Possibilities of Our Grandchildren*, edited by Lorenzo Pecchi and Gustavo Piga. Cambridge, MA: MIT Press, 2008, 20-1.

Krugman, Paul. "Paul Krugman Reviews 'The Rise and Fall of American Grown' by Robert J. Gordon." *New York Times*, January 25, 2016, https://www.nytimes.com/2016/01/31/books/review/the-powers-that-were.html.

Kumar, Rahul. "Risking and Wronging." *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 27–51.

Kurtz, Annalyn. "World's Shortest Work Weeks." *CNN Money*, July 10, 2013. https://money.cnn.com/gallery/news/economy/2013/07/10/worlds-shortest-work-weeks/5.html.

Mishel, Lawrence, and Josh Bivens. "The Zombie Robot Argument Lurches On." *Economic Policy Institute*, May 24, 2017. http://www.epi.org/files/pdf/126750.pdf.

Panozzo, Chantal. "Living in Switzerland Ruined Me for America and Its Lousy Work Culture." *Vox*, February 1, 1016. https://www.vox.com/2015/7/21/8974435/switzerland-work-life-balance.

Petroff, Alanna. "German Workers Win Right to 28-Hour Work Week." *CNN Money*, February 7, 2018. https://money.cnn.com/2018/02/07/news/economy/germany-28-hour-work-week/index.html.

Pettit, Philip. "A Republican Right to Basic Income." *Basic Income Studies* 2, no. 2 (2007), 1-8.

Rawls, John. *Justice as Fairness: A Restatement.* Cambridge, MA: Harvard University Press, 2001.

Rawls, John. *A Theory of Justice.* Cambridge, MA: Harvard University Press, 1971.

Samuelson, Paul. "Bergsonian Welfare Functions." In *Economic Welfare and the Economics of Soviet Socialism*, edited by Steven Rosefielde. Cambridge: Cambridge University Press, 1981.

Savage, Maddy. "What Really Happened When the Swedes Tried Six-Hour Days?" *BBC News*, February 8, 2017. https://www.bbc.com/news/business-38843341.

Scanlon, T. M. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.

Solow, Robert. "We'd better watch out", *New York Times Book Review*, July 12 (1987).

Solow, Robert. "Whose Grandchildren?" In *Revisiting Keynes: Economic Possibilities of Our Grandchildren*, edited by Lorenzo Pecchi and Gustavo Piga. Cambridge, MA: MIT Press, 2008, 87-93.

Stokey, Edith, and Richard Zeckhauser. *A Primer for Policy Analysis* (New York: W. W. Norton and Co., 1978).

Summers, Larry. "Economic Possibilities for Our Children." *NBER Reporter*, no. 4 (2013), 1-6.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Knopf, 2017.

Turner, Karen. "Amazon Is Piloting Teams with a 30-Hour Workweek." *Washington Post*, August 26, 2016. https://www.washingtonpost.com/news/the-switch/wp/2016/08/26/amazon-is-piloting-teams-with-a-30-hour-work-week/?utm_term=.c96516b65c36.

Van Parijs, Philippe. *Real Freedom for All: What (If Anything) Can Justify Capitalism?* Oxford University Press, 1998.

Wray, L. Randall. *Modern Money Theory: A Primer on Macroeconomics for Sovereign Monetary Systems*. 2nd ed. New York: Palgrave Macmillan, 2015.

7

Autonomous Weapons and the Ethics of Artificial Intelligence

Peter Asaro

While "killer robots" have long been a staple of science fiction dystopias, they also represent a critical and central issue in the ethics of artificial intelligence (AI). More technically speaking, autonomous weapons are a real and emerging technology that have the potential to radically transform warfare, policing, and how we understand human rights in relation to the operations of machines, algorithms, and AI. The issues raised by giving machines the capability and, more important, the *authority* to kill human beings raises a range of ethical as well as legal, social, and political issues. Many of these issues are of critical importance even if we consider only simple forms of automation, or *artificial stupidity*. Other issues arise if we consider the difficulty of properly gauging the capabilities and reliability of increasingly sophisticated forms of AI. And yet other issues arise if we consider the remote advent of some form of an artificial general intelligence (AGI), human-like AI, or superintelligence. Because the issues raised by simple autonomous weapons are the most urgent, I will focus on these. But I will also consider some of the issues raised by increasingly capable systems, and reflect on the implications of highly capable future AI.

In considering a few of the most significant issues raised by autonomous weapons, I will seek to articulate them according to some major philosophical approaches to ethics. As such, I will not endorse any particular approach. Roughly speaking, my approach is that where there is broad agreement that moral rights and duties exist and are clear, they provide reasons that are more compelling than utilitarian reasons, while utilitarian reasons are useful in the absence of clear moral duties and rights. Further, I believe that moral virtues and sentiments reflect psychological and cultural norms and preferences and often function as heuristics in moral reasoning, especially when one must choose between competing duties and values and when one reflects on the implications for one's own moral character when taking an action. While no single moral theory alone can fully explain our views of autonomous weapons, each of the leading Western moral theories—deontological and consequentialist—points to the immorality of autonomous weapons in its own way, and taken together the

Peter Asaro, *Autonomous Weapons and the Ethics of Artificial Intelligence* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0008

leading moral theories present a clear case that building and using autonomous weapons, and permitting or authorizing autonomous violence, is morally wrong.

7.1. Defining Autonomous Weapons

Let us start by considering a simple working definition of what constitutes an autonomous weapon. Modern weapons and weapons platforms utilize a great deal of automation in the operation of various functions of the system and at various levels of control. For instance, a guided missile contains feedback regulators over the direction of thrust so as to direct the missile toward some target. In a sense, the system is guided toward a goal or target according to some control mechanism and a sensor (heat, electromagnetic, or optical), or some coordinate guidance system, such as satellite-based global-positioning systems (GPS). The missile is assigned its specific goal by a human operator, who locks it on target using a laser or radar system or provides GPS coordinates. The automation then follows some set of control parameters such that it advances toward its goal and adjusts its controls to keep the missile on course until it strikes its designated target.[1]

Remotely piloted vehicles, better known as drones, also contain a great deal of automation. With what amounts to sophisticated autopilot systems, these drones can be given preset flight paths, or a series of GPS way-points, and automatically fly to them. Along the way they make numerous automated flight-surface and throttle-control adjustments to compensate for wind, thermals, aberrant sensor data, and more. Some advanced drones are also capable of automated takeoff and landing and automated aerial refueling, and researchers are regularly achieving increasingly sophisticated automated maneuvers. Of course, some drones carry the same missiles and bombs found on other military aircraft, which can themselves have complex automated guidance systems. They are thus considered weapons platforms for these weapons, containing multiple levels of control.

However, neither guided missiles nor remotely operated drones are autonomous weapons in the relevant technical sense that concerns us. Following the working definition offered by the International Committee of the Red Cross (ICRC), an autonomous weapon system is any system that automates the critical functions of targeting and engaging a weapon.[2] This means that the targeting and use of force must be automated for the system to be considered an autonomous weapon. Another way of looking at it is that autonomous weapons systems *lack meaningful human control over the critical functions of targeting and use of force*.

Whether such systems already exist depends on how one interprets *meaningful human control*, a topic to which I will return. A variety of existing systems use some form of automated targeting. In particular there are mines and

booby-traps that are victim-activated by pressure or proximity sensors, or even acoustic sensors; sentry guns that are similarly victim-activated by motion sensors; loitering munitions that search out specific radar signals over large areas; and a host of projectile-intercept systems that automatically track and target incoming missiles and mortars and shoot them down. Most missile defense systems[3] are autonomous for only a few seconds at a time, however, and remain under the direct control of humans who can observe their operation, scrutinize their targets, and deactivate them at any moment. So they could be argued to be under meaningful human control (or *not*, if one further requires strict individual target authorization), while wide-ranging loitering munitions, mines, and sentry guns systems are more problematic. This issue partly turns on the ethical aspects of such targeting, so we will return to it later.

In the case of remotely operated drones, a human (albeit from very far away) interprets the imagery data, identifies potential targets, verifies and selects a target, and then aims and engages a weapon. The drone operator may also consult human intelligence analysts and lawyers and seek authorization from superior officers. Typically, such armed drones employ a laser that they shine on the target—and automation helps to ensure that it stays on that target—which the sensors on the drone's missile can use to guide it on a path to the designated target. Thus, while such systems use automation, they are not autonomous weapons.

If, however, the drone utilized automated software to scan through its video feeds and sensor data, automatically identified targets, and then selected and fired on those targets, all without human intervention, supervision, or control, then that would clearly be an autonomous weapon. From a causal or functional perspective, this may not seem like a large or significant difference—just a bit more automation, or automation in a different stage of the operation of the system. But, of course, the difference is one with ethical, as well as legal and political, significance. While there are different ways to frame the operations of these systems in moral terms, I will argue that automating the targeting of weapons and the use of violent force necessarily has ethical and moral significance and should be recognized as such. Further, I will argue that as the development of such systems has become technically feasible, we should recognize existing moral and legal principles and establish new norms that clearly prohibit delegating the authority to kill to machines.

Another way of looking at such a norm is as a positive obligation to ensure meaningful human control over the use of weapons. Of course, the concept of "meaningful human control" will require some articulation, but the basic idea is that morally and legally responsible human agents must retain control over the functions of any system that directs and releases violent force. But it is also true that a robust conception of meaningful human control could find useful

applications to other problems in AI ethics. Such applications include the relationship between self-driving cars and their occupants or operators and in the application of algorithms to decisions with the potential to deny or deprive humans of their fundamental rights, from access to medical care, credit, educational and employment opportunities to exercising their economic, cultural, and political rights. In all of these cases, there is a potential to delegate an authority to a machine that might directly impact the moral and legal rights of a human person.

7.2. The Moral Problems Raised by Autonomous Weapons

In providing a moral analysis, it will also be helpful to lay out the various types of issues that have been raised as problems with autonomous weapons. These represent a range of different concerns and can potentially be characterized differently under different moral theories. Yet each set of concerns also lends itself to one or more moral approaches. The concerns can be grouped together into some broader categories: harms to civilians, arms races and international instability, intrinsic unpredictability, hacking and cybersecurity threats, a new type of weapon of mass destruction, threats to responsibility and accountability, and threats to human rights and dignity. Along the way we will also consider arguments that autonomous weapons and their use might be morally superior to human-controlled weapons.

7.2.1. Harms to Civilians

By far the most commonly expressed concerns around autonomous weapons are that they will kill innocent civilians and destroy civilian infrastructure. Such a concern may seem quite simple and straightforward, but there are different ways to characterize this worry. Following the formulations of international humanitarian law, which requires military attacks to be discriminate and proportionate, one could argue that autonomous weapons will be indiscriminate in their targeting, failing to distinguish civilians from combatants. One could also argue that autonomous weapons might make disproportionate attacks, killing many civilians for a relatively low-value military objective. One could alternatively argue that autonomous weapons would lack aspects of human psychology that might make them more humane in warfare. They might thus be far more aggressive or fail to show any compassion in situations where a human might be merciful. Worse, autonomous weapons could be easily designed, altered, or manipulated to purposely harm civilians (i.e., given such a goal either explicitly

or implicitly). Despots and tyrants might turn such weapons against their own people or apply them to genocidal ends, or terrorists might use them to attack civilians. Despite the various ways autonomous weapons might cause negative impacts on civilians, it is possible to group these concerns together.

On initial consideration, this looks like a consequentialist concern: there will be significant negative consequences for civilians if autonomous weapons are deployed. Of course, it could also be viewed as violation of the rights of those individuals and thus a deontological concern, which we will consider shortly. But in a consequentialist evaluation, whether the use of autonomous weapons is morally good or bad depends on numerous empirical facts about the actual impacts and the probabilities of those impacts, which are difficult to assess before such weapons are used. Indeed the proponents of developing autonomous weapons often argue on the same consequentialist grounds that autonomous weapons could be designed to be far better than humans at making targeting decisions and conducting attacks, thus reducing the risks of harm to civilians.[4] From an engineering perspective, these negative consequences can also be viewed as risks, and systems can be designed to try to minimize or eliminate such risks. This kind of framing sets up the design of autonomous weapons as a form of safety design: maximize killing "bad guys" while minimizing the killing of "good guys." But while this might be reasonable as an argument for reducing unintended and undesired killing, it does not fully address the morality of the intended automated killing.

Upon further reflection, one can also look at this as a deontological issue: the negative impacts on civilians in these situations will be death, grave injuries, trauma, and displacement. While these are, of course, bad consequences, they are also the deprivation of fundamental human rights—the rights to life, bodily integrity, and dignity. Under this view, we have a moral duty to respect the rights of others and to treat them as ends in themselves. But autonomous weapons could prevent us from performing our duty to respect others in war. Deontological ethics does not completely prohibit killing, particularly in war, but it does require that there be a valid justification, such as self-defense, for killing. Similarly international law permits the unintentional killing of civilians in war, provided there was a lawful (justified) military objective for an attack. But it is less clear what counts as "intentional" or "unintentional" when it comes to autonomous systems; operators may have an idea of what an autonomous system will do, but it may do many things they did not specifically consider and kill people the operators did not intend to kill. In most circumstances we view such situations as accidents and hold people morally responsible only if they were acting in a reckless or negligent manner. Autonomous weapons, particularly those involving advanced AI, can serve to cloud the moral issue, however, insofar as the operator of a weapon might believe that the system is designed to attack only valid military

targets and to avoid and protect civilians. If this is a reasonable belief, then we might be inclined to attribute any harms to civilians as simple accidents or technical failures, and not the moral responsibility of the operator.

More critically, there is a question as to whether the use of an autonomous weapon might be a means of fulfilling one's moral duty to respect the rights of others, or actually precludes the ability to respect those rights. As Sparrow[5] demonstrates, it could be argued that if there is a weapon system designed to protect civilians, and that actually works in that way, we may have a moral obligation to use it. While I see the consequentialist side of this argument, as discussed with respect to Arkin,[6] I do not see the deontological side of it. In particular, in order to fulfill our duty to respect the human dignity of others, I believe we are required to recognize them as human and to consider them as such when making the decision that it is justified to kill them or put them at risk of death. Insofar as this consideration is not actually taken by the operators of the weapon, but they instead rely upon an automated process, then they are not really fulfilling this duty. Indeed they do not necessarily even think about the individuals that may be killed, much less regard them as persons. And as we will see, neither does the automated system. This has implications for both the respect given to human rights and dignity and to how we ascribe moral responsibility and legal accountability.

7.2.2. Arms Races, Rapid Proliferation, and Instability

Another broad range of issues concerns the impact of the introduction of autonomous weapons in the context of international relations. Insofar as these weapons are seen as high-tech and prestigious, as well as providing tactical or strategic advantages over the capabilities of adversaries, or serve as an effective deterrent, there will be strong incentives for countries to develop or obtain such weapons. The same logic, of course, applies to their adversaries and competitors. This is the logical foundation of a competitive arms race wherein rivals expend large amounts of resources in an effort to gain military advantage over their competitors.[7] Apart from being an expensive use, or waste, of economic, intellectual, and natural resources, such arms races are tied to political and military instability.[8] Since significant military buildups and strategic advantages are viewed as threatening to neighbors and adversaries, some states may consider preempting such advantages rather than allowing them to develop. As such, arms races can raise tensions and create instability. Having access to new high-tech weapons, especially ones untested in real conflicts, can also give leaders a sense of having superior military capabilities, which in turn makes them more inclined to initiate or escalate military actions—whether or not their confidence is warranted. And insofar as the weapons themselves may behave or perform in unexpected

ways due to AI, they become less predictable as threats by adversaries, leading to greater instability. Such arms races and rivalries can operate at regional levels between neighboring states, or at global levels between superpowers and groups of aligned states.

Again, from a consequentialist perspective, whether such arms races are good depends on one's evaluation of the outcomes, as well as how such rivalries play out. Some might argue that the Cold War was an arms race that ended in a stalemate of sorts, or detente, and was preferable to war. Others might argue that the Cold War led to numerous proxy wars and that there were many other ways this rivalry might have played out, short of war, that did not require massive investments in nuclear arms and their delivery systems and that could have had much better outcomes. If we examine the wider set of outcomes and their probabilities, including the possibility of nuclear war and its history of near-misses, what we see is that instability is itself undesirable in international relations as it is a leading factor in many conflicts, including "low-intensity" and proxy conflicts. Instability makes it harder to predict how one's adversaries might act, as well as determine the severity of the threat, which makes it more likely that one will take a defensive or proactive stance. Insofar as such buildups are viewed as intrinsically or implicitly hostile, the very existence of an arms race is a manifestation of hostilities between adversaries. In short, such instability increases the probability of violent conflicts occurring, which is bad on utilitarian grounds.

Closely related to concerns over arms races between states is the concern that autonomous weapons will proliferate rapidly. From the perspective of great power states, there is a concern that smaller states might rapidly acquire significant military capabilities to challenge larger wealthier militaries. Because autonomous weapons do not require the industrial and technical sophistication that nuclear weapons do, and simple ones can even be built with off-the-shelf technologies, many states will acquire them rapidly. We have seen this already with surveillance drones, and now increasingly with armed drones. As we have also seen with drone technology, autonomous weapons, including the advanced, sophisticated, and hardened types that major militaries would develop, would also find their way to nonstate actors or terrorists or even be acquired by police forces. At least with the more sophisticated versions, they are likely to be developed only by those with significant resources, and thus a strong stigmatizing norm against autonomous weapons could prevent such systems from being developed or produced in significant numbers.

One could also consider the wider implications of the drain on resources such arms races will entail.[9] For all the discussion of "AI for good" and "socially beneficial AI" in the AI community, if many or most of the best AI engineers end up working on expensive military AI and robotics projects, they will not be working

on those socially beneficial projects. It is difficult to measure the value of such missed opportunities, but it would be substantial.[10]

7.2.3. Unpredictability

Another fear is that autonomous weapons could simply go out of control and do things that are completely unintended or highly unpredictable. While armed conflict is always unpredictable, such systems could add a whole new level of unpredictability. On the one hand, there is the possibility of such systems initiating or escalating a conflict without any human political or military decision to do so. While this can sometimes happen due to the unauthorized actions or mistakes of military personnel, humans are capable of recognizing the larger implications of their actions and can seek confirmation from superiors, while automated systems are not capable of this.[11]

The operator of an autonomous system may have a general idea of what the system is supposed to do, and may further have operational experience of how it operates in various specific contexts. But insofar as autonomous weapons are designed to operate over large geographic areas and time frames, and given that the possible interactions with the environment it may have grow exponentially, it will become increasingly difficult for even well-trained operators to reliably predict what a system will actually do once deployed. Testing and reliability metrics can offer confidence to operators only when systems are deployed in situations and contexts that match those under which it has been previously tested, while increasing ranges and time frames imply that operators are less aware of the specific characteristics of the environment the system will encounter.

Further, there is much interest in developing large fleets and swarms of autonomous weapons systems. Such large ensembles of autonomous systems, even relatively simple ones, are known to be intrinsically unpredictable, from a mathematical perspective. But even a small number of autonomous systems interacting with each other, when we know how only some are programmed, are unpredictable because we do not know how an adversary's systems are programmed or what the net result of interactions between them will be. This issue is similar to that of various computer trading systems, whether for pricing products for online markets like Amazon or for trading stocks. Both have manifested unexpected positive feedback loops resulting in million-dollar books being listed for sale on Amazon and in major trading market crashes, such as the one at the New York Stock Exchange in 2010, called a flash-crash, that lost 9% of the market's value in just a few a minutes.[12] However carefully programmed and tested autonomous weapons are, such catastrophic incidents will remain highly probable or inevitable if large numbers of autonomous systems are deployed. The consequences,

however, could be far worse if those systems are controlling weapons instead of trading stocks or selling books.

7.2.4. Lowering Thresholds of Conflict, Unintended Conflicts, and Unattributable Attacks

Another set of concerns around autonomous weapons is that they will dramatically shift how political leaders think about armed conflict, how they make decisions about the use of military force, and even how military leaders make strategic decisions. Because autonomous weapons promise to deliver military goals without putting human soldiers at risk, such weapons could lower the political thresholds for going to war. As we have already seen with armed drones, which create the possibility of military interventions without risks to either pilots or special forces commandos, leaders may be more likely to choose a military option offered by autonomous weapons when other options are too politically risky. If such situations are common, then the result will be more military operations rather than seeking political solutions.

There is also a set of concerns around the possibility of autonomous weapons acting or reacting in ways that initiate or escalate a conflict without any human political or military decisions. Imagine a border patrolled by autonomous combat aircraft from neighboring countries. One might be blown off course and into the airspace of the other, which could automatically initiate an attack; the other could return fire, both could call in additional units, and very quickly a conflict could be initiated before humans were even alerted. Similarly, a low-level military operation, such as a patrol, could rapidly escalate due to a series of automated decisions into a major engagement, which could in turn escalate the nature of the overall conflict, or lead to the commitment of greater resources and more extreme forms of violence, or lead to the involvement of previously neutral parties. This concern is also related to the unpredictability concern of the previous section.

Autonomous weapons, even more than remote-operated weapons, offer the possibility of unattributable attacks. This is a phenomenon usually discussed in cyberwarfare, where it can be very difficult to determine the source of an attack with enough certainty to justify economic, political, or military retaliation. But insofar as there is plausible deniability, or genuine uncertainty, as to who built and deployed a weapon, and its purpose, it will be difficult to definitively attribute an attack to its author. This could lead to widespread use of assassination by states against perceived foes, or seemingly random and chaotic attacks meant to destabilize and confuse civilians or political leaders. The possibility that systems, including those with known owners, could be hacked and hijacked and

turned against third parties could also cause significant problems, to which we now turn.

7.2.5. Vulnerabilities to Hacking, Spoofing, and Cyberattacks

It is possible to create autonomous weapons that do not use programmed computers; we could even consider landmines as the "stupidest autonomous weapons" on the basis of their lack of discerning sensors or computational functions. However, it is much more likely that we will see computational technologies involved in the decision processes of most autonomous weapons, as well as a variety of sophisticated sensors providing inputs to those decision functions. And given the nature of armed conflict, it is very likely that adversaries will attempt to interfere with autonomous weapons directly through cyberattacks that impair, disable, or take control of those systems, or indirectly by fooling or "spoofing" those systems through their sensors and what is known about how they process information.

Spoofing is a form of tricking an automated system to do what you want it to by manipulating its sensor data. One could do this by attacking its sensors or simply manipulating what those sensors capture. A well-known example of this is spoofing GPS geolocation sensors. These sensors respond to signals from GPS satellites in space and compute their location from the signals of multiple satellites. It is possible to bombard these sensors with signals that imitate the satellite signals but are much stronger. If, for example, an autonomous drone aircraft is attempting to fly to a certain coordinate, it is possible to force it to fly wherever you want by systematically manipulating its GPS inputs.[13] It is not unreasonable that this and many other means might be deployed to spoof autonomous weapons, including baiting them to attack the wrong targets, expend their ammunition, or even turn them against civilians or the military that fielded them.

Indeed recent research in machine learning has demonstrated that because the data spaces over which deep-learning algorithms learn is so vast, it is possible to develop what are called generative adversarial neural networks which can systematically deceive a trained neural network so as to trigger any desired output. Moreover it can do this with, for example, visual images that appear to the casual human observer to be identical to images that normally have a very different output. For example, two images of a "Stop" sign might appear identical to human observers, yet one could cause a self-driving car to stop as it should, while the other could be designed by an adversarial network that figured out a few select elements of the image that, when altered, will trigger the neural network to instead recognize this as a "Speed Limit 55" sign. This is a major and fundamental problem within machine learning, with no apparent solution. As long

as it remains unsolved, any autonomous weapon that employs such machine-learning techniques would be susceptible to manipulation, including making enemy combatants look like civilians, and friendly forces look like threatening adversaries.

Of course, because autonomous weapons will be primarily computer-controlled, and likely networked, they will be subject to most of the same vulnerabilities currently faced by computers and computer networks; namely, hackers will be able to launch cyberattacks against them and potentially gain control of them. While that is possible to some extent with advanced weapons systems that employ computer controls, insofar as those systems require human operators to engage the weapons or pull the trigger, these functions cannot be commandeered by hackers. Autonomous weapons will extend the power of hackers and the kinds of effects they can have.

As discussed earlier, it can often be difficult to attribute cyberattacks to their source. Insofar as unattributable cyberattacks can gain control of weapons systems, then attacks from those weapons systems will also be unattributable. Thus hackers could commandeer the weapons of one country and use them against another country—and it could turn out that neither country is able to determine the source of the attack. Alternatively, one country could simply claim that its systems had been commandeered and launch an attack. It will become increasingly uncertain and more difficult to establish attribution, resulting in more plausible denials and highly unstable political situations, leading to greater international instability.

7.2.6. A New Kind of Weapon of Mass Destruction

Another concern raised by autonomous weapons is that they could constitute a new form of a weapon of mass destruction (WMD). Historically this term referred to nuclear weapons and later to chemical and biological weapons that could have massively devastating effects from a single use, similar to that of a nuclear weapon. But a more technical way of looking at or defining a WMD would be to say that it is a weapon that allows a single individual or small group to cause mass casualties. Conventional guns and bombs can cause mass casualties, but only to a degree far lower than what a nuclear bomb or the poisoning of a water supply for a major city might. Because autonomous weapons do not need individual operators, it seems likely that a single individual or small group will be able to deploy vast fleets or mass swarms of such weapons. Unlike the indiscriminate nature of previous WMDs, such swarms of autonomous weapons could be designed to be highly discriminate, for example, killing everyone over a certain height or whose face matches a face in a database. The point is that small groups

could unleash mass devastation. This is worrying because it could further empower terrorists, tyrants, and others who would wreak such devastation and sow chaos and fear in order to enhance their own position or power, thereby serving to destabilize the peace and security of the world.

7.3. Arguments for the Moral Desirability of Autonomous Weapons

There are those who have argued that autonomous weapons are not only morally permissible but are morally desirable or even morally required. While I disagree with this view, it is important to understand the basis of this argument in terms of moral reasoning. That basis is ultimately a consequentialist, that is, utilitarian, one, and while compelling when considered as a singular decision regarding immediate consequences made in isolation, it fails to take seriously any of the broader consequences of permitting autonomous weapons.

The basic argument, as articulated by the roboticist Ron Arkin,[14] is that many of the civilian casualties in war are the result of human errors: soldiers making lethal decisions when they are exhausted, afraid, angry, or even vengeful. The "plight of the non-combatant," as Arkin[15] puts it, is not simply being at the wrong place at the wrong time and getting caught in the crossfire, but facing death owing to the human failings of the soldiers who are entrusted to pull the triggers. If this is indeed the plight of noncombatants in warfare, then introducing automation that could "correct" those mistakes would greatly reduce civilian casualties. Given that there is an obligation to protect civilians from combat, it follows that it is desirable and perhaps even morally required to automate lethal decisions so as to eliminate such human errors.

Apart from the lack of evidence for its empirical claims,[16] this argument takes a very narrow view of morality with respect to killing. It imagines that we can directly substitute a human targeting decision with a computational process that will perform better than a human with respect to identifying civilians. Even if such a computational identification system existed, this by itself does not seem to require the elimination of the human decision-maker. The computational system could be an advisory or recommendation system, allowing the human to make the decision and take action, but also provide warnings about potential errors in judgment or the risks of alternative actions. We could even go further and design the system to not permit humans to target civilians or put them at risk at all. Like an automated safety that prevents a weapon from firing on civilians, Arkin et al.[17] call this mechanism "the ethical governor" after the steam-engine governor of James Watt. Again, the ability of such computational processes to increase accuracy and precision in distinguishing civilians or to better predict the risks of

certain attacks to those civilians does not directly argue in favor of completely eliminating the human element. The only reasons in favor of that are reasons of efficiency or expediency and perhaps the military advantages of deciding and thus acting more quickly. Similarly one could argue that, unlike machines, humans are too slow and suffer from fatigue or psychological pressures. Those may be good reasons for accounting for human failings and in certain situations when response time is critical, such as in the heat of battle. But in terms of overall military operations, live combat is a relative rarity, time is not always of the essence, and sometimes patience is rewarded with more favorable conditions for completing a mission or reducing risks to civilians. Indeed much thinking around nuclear arms control is concerned with increasing the amount of time to make any potential launch decision.

While these arguments do not lead necessarily to the conclusion that humans will always make better decisions than machines, they do significantly weaken the intuition that replacing human decision with high-performing machine decisions will necessarily be good or even better than human decisions, or decisions reached by humans augmented with advisory information systems. Again, the basic structure of Arkin's argument, while appealing to the rights of civilians and duties of soldiers to protect civilians, takes a consequentialist form in which whatever reduces civilian casualties the most is the most desirable approach. But for this to be true, one must consider a broader view of the consequences of permitting autonomous targeting and the use of violent force without meaningful human control.

From the perspective of international relations and global security, the push toward greater autonomy in weapons has been regarded by some to be a revolution in military affairs with the potential to change the nature of warfare and the balance of power between those who have such weapons and those who do not. As such, there are clear risks to regional and global political stability and precarious balances of military power, owing to potential arms races and proliferation in the domain of autonomous weapons, as described earlier. Moreover, as these weapons go into mass production, their cost will go down and their availability will increase, making them much easier to obtain by nonstate actors and terrorist organizations.

The kinds of potential problems, risks, and harms, described in more detail in the previous sections, are primarily consequentialist in nature. As such they depend on how the development, deployment, and use of autonomous weapons actually play out in the real world. At this point, we can identify the most likely uses and risks posed by the technology. There are, of course, other possibilities for how the technology might unfold, paths it might take that we do not expect. Despite this, it seems unlikely that those paths would lead to more restrained development, more restrained uses of force, or lower risks to civilians

and combatants. What seems far more likely is that the development and use of autonomous weapons will lead to more conflicts of increasing intensity and an overall rise in political instability within countries and internationally. Thus even if we were to accept the utilitarian advantages of increased targeting precision of autonomous weapons, the overall consequences for civilians in war, and humanity in general, could be negative, and appears very likely to be so. But our ethical and moral considerations of autonomous weapons need not be limited to consequentialist analysis or assessments of the value and likelihood of various possible outcomes. Rather we can look to the impact the development of autonomy in weapons systems will have on key moral principles of responsibility and accountability, human rights and human dignity, to which we now turn.

7.3.1. Threats to Responsibility and Accountability

Unlike previous technological advances, the advance of autonomy in machines presents some unique moral challenges. This is because machine autonomy intercedes on human agency, both redistributing it and rearranging it and in some ways confounding the norms we have long relied upon for ascribing moral responsibility and holding people and institutions accountable.

In particular, the delegation of targeting decisions to machines poses a specific threat to ascribing responsibility to the operators or commanders of an autonomous machine. Generally, when a human is in control of a weapon system and directs that system at a target and engages it, we expect that the operator has an intention to destroy or disable the target. The target must be recognized as valid or legal, destroying it must be recognized as fulfilling a military necessity, and the use of force in the situation must be morally justified, legitimate, or permissible.

However, if the operator has delegated the targeting decision to an autonomous function of a machine, then the targets are determined by algorithms applied to data. While the operator may have a general intention in mind, such as "Destroy enemy vehicles in this area," unless the operator inspects and confirms each of the targets selected by the automated process, the operator does not know if each is, in fact and in this particular circumstance, a lawful target. As such, neither the operator nor the system designers have access to the justification for designating a target as lawful. And if the system were to make mistakes, these might not be viewed as culpable crimes but rather as mistakes or simply technical errors.

Moreover, the system itself, and its algorithms, are not legal or moral agents that can be held morally responsible or legally accountable for their choices and actions. While there is a certain intention behind the design of an algorithm, many assumptions must be made about the context and circumstances in which

the algorithm will operate and the kinds of sensor data it will receive. As such, algorithm designers are not making judgments based on the actual circumstances and situation in which those decisions will be made; they are simply crafting clever rules that they expect will approximate such judgments given various assumptions. While there may be a certain degree of accountability for the designers of systems, particularly if they are negligent in their designs, it is quite natural to excuse mistakes owing to unforeseen circumstances.

Researchers in machine ethics have suggested that we can model or simulate legal and moral reasoning in a machine. But even if we try to represent international humanitarian law in a computational system,[18] and provide a means to reason out whether a particular target is lawful or not, this kind of simulation is not sufficient to justify killing. If the system were to kill a civilian, how would we hold it accountable or responsible for that death? We might be able to ask for its justification, the chain of reasoning that led it to make the incorrect decision. Or we might be able to diagnose the failure in its sensors or logic or the features in the environment that led to the error. And we might even be able to correct these failures. But the system itself would not be responsible and could not be punished.[19] And since we cannot really hold the operators or designers responsible either,[20] there would seem to be a responsibility gap that has been created by the introduction of the autonomous system.

In civilian cases, there is a body of liability and tort law in place to address the nature of responsibility in unintentional and accidental harms.[21] However, in warfare such laws do not apply to combatants. Some have argued that there are indeed war torts,[22] yet this applies mainly to the liability of military suppliers and subcontractors to the military, not the liability of soldiers to their accidental victims.

Others have argued that states will always be responsible for the armed forces they command, and thus for any autonomous weapons they deploy. However, there is a very different psychological process involved in soldiers who are making a judgment to use lethal force for which they will be responsible, and deciding to deploy a system for which one does not expect to be held accountable. As we have seen in other areas where automation has been deployed, the lack of potential responsibility creates greater levels of risk-taking and recklessness. Consider the vast array of financial instruments designed to limit risk and liability, and the high-risk markets they have spawned.[23]

Since machines and automated functions are not moral and legal agents, it is inappropriate to delegate moral and legal authorities to such systems.[24] In the case of autonomous weapons, it is immoral, and should be illegal, to delegate to such systems the authority to kill or to select and engage targets with violent force. The legal consequence of such delegation is to create the responsibility gap, which undermines the moral and legal responsibility of the individuals involved

in armed conflict. And further, insofar as it becomes more difficult to apply international law to individuals, it also serves to undermine the international law framework itself.

From a deontological perspective, one can delegate the performance of certain obligations to other moral agents who take responsibility for fulfilling those obligations. But those agents must be moral agents capable of taking that responsibility. It is irresponsible, and thus immoral, to delegate obligations to entities that cannot take on those responsibilities.[25] From a consequentialist perspective, such delegation might look appealing if we believe that the amoral agent might act to increase overall utility. However, in calculating the balance of utility, it should be noted that there is a moral hazard, itself a negative effect, in allowing individuals to wrongfully delegate their obligations because they will be less likely to take their obligations seriously or feel responsible for their moral failings or act to fulfill their obligations. Similarly, from a legal perspective, the inability to hold individuals accountable or responsible for their actions or failures to fulfill their obligations makes legal enforcement difficult or impossible. This, in turn, is likely to lead to the flouting of those laws, as well as a more general loss of respect for all laws and the legal framework itself. Thus the abrogation of duties through wrongful delegation both is deontologically wrong and has serious negative consequences.

The other way to view this moral requirement is that humans need to maintain control over weapons systems to the extent that they can ensure the targeting of humans is lawful and the use of violent force against them is justified. Another way of stating this is that all weapons systems require *meaningful human control*. Before considering just what this might mean, we turn first to a discussion of the nature of human rights and human dignity, which are threatened by the lack of meaningful human control of autonomous weapons.

7.3.2. Threats to Human Rights and Human Dignity

The issue of autonomous weapons was first raised at the United Nations by Christof Heyns[26] in his report to the Human Rights Council as special rapporteur on arbitrary summary and extrajudicial execution. He argued subsequently that the fundamental ethical question is one of rights and dignity, and the ICRC has since reached the same conclusion.[27] Many states have also acknowledged the difference between consequentialist arguments for and against LAWS, and deontological arguments over the impact on rights and dignity. Of course, there is also a great deal of misunderstanding regarding deontological ethics and arguments, which some people find intuitively compelling and others do not.

The right to life is widely recognized as a fundamental human right, the loss of which is irrevocable and upon which nearly all other rights depend. One cannot exercise one's right of free speech if one is dead, nor can one's life be restored if it is taken in error. As such, the right to life is highly valuable, and any decision to deprive someone of life is of great significance and requires compelling justification. While accidents deprive people of their lives with some frequency, these are not intentional acts and thus we do not expect them to have justification. If we allow autonomous systems to target and engage violent and lethal force, however, we must ensure that the intentional killing that results is legally and morally justified. And, as discussed in the previous section, insofar as artificial systems are not capable of legal and moral agency or of appropriate legal and moral deliberation, they cannot understand whether a particular killing is justified, nor can they be held responsible for such a judgement.

Human dignity is a concept that often appears in discussions of human rights but is rarely considered in detail in arms control. It is sometimes described as a sort of property that attaches to persons and that can be stripped from them. But this metaphor captures only a part of what constitutes human dignity and does little to help us understand its importance in armed conflict. Indeed it is often said that there can be little dignity in war, or in dying in war, and certainly the manner in which many people are killed in wars—by flames, explosions, shrapnel, bullets, and so on—is lacking in dignity. But this is not what is really meant by human dignity or what it means to respect it. It is not the physical means of death that determines whether a death is dignified, any more than the manner of death justifies whether the death is lawful or moral. The human right to dignity, much like the human right to life, inheres not in a property and its loss or the physical-causal means of its loss but in the reasons for that right being violated or overridden. And the fundamental right to dignity is the right to be recognized as a human and accorded the respect and rights of all humans.

The human right to dignity, like the human right to life, is an intrinsic right that is realized in relations between humans—and duties of humans to respect the rights to dignity and life of other humans. Such rights are never lost, merely overridden by the rights of others, for example, to self-defense. Accordingly, there is no "right to kill" or "license to kill," even in war. Rather, the right to defend oneself individually, or to defend one's nation collectively, is recognized as the right to life of one party overriding the same rights of another. It is also recognized as limited and requiring the proper relation between individuals. Thus soldiers can be murdered by civilians in war, just as soldiers can murder civilians, or even fellow soldiers, in war, yet the killing of enemy combatants is not considered murder. But it is only when the proper relation exists, that is, that each are enemy combatants in a state of war, that killing is morally and legally acceptable. And further, establishing that relationship depends upon *reasons*, primarily

the right to self-defense but also satisfying the conditions that limit killing in war: discrimination to ensure that only those in the proper relation are killed, proportionality to ensure that killing is not disproportionate to its justification, and military necessity to ensure that the killing is really necessary for the ostensive purposes of collective self-defense. When killing lacks these properties it can be considered a war crime, or it can be an arbitrary, summary, or extrajudicial execution in which there are not sufficient moral or legal reasons, reasoning, or legal process to justify the killing. For killing to be nonarbitrary, a morally responsible agent must have legitimate reasons for depriving someone of life.

When it comes to the question of autonomous weapons, it may seem easy to argue that it does not matter how one is killed in war. But clearly it does, from both a moral and a legal perspective. What, then, is required to ensure that killing in war does not violate human dignity? As Heyns[28] has argued, it must not be arbitrary, which is also to say that it must be justified. The question for autonomous weapons is whether a calculated machine decision—using computations based in sensor data—can understand and act on legitimate reasons for killing.

Elsewhere[29] I have argued that computational systems are not moral and legal agents and thus cannot legitimately determine when it is appropriate to take human life. While computational systems may be able to accurately and reliably apply a computational rule to a set of data, in so doing they are not thereby respecting human dignity. In order to make a moral judgment to take a life, while respecting human dignity, it is minimally required that a moral agent can (1) recognize a human being as a human, not just distinct from other types of objects and things but as a being with rights that deserve respect; (2) understand the value of life and the significance of its loss; and (3) reflect upon the reasons for taking life and reach a rational conclusion that killing is justified in a particular situation. Currently only humans are capable of meeting these criteria, which is why it is morally and legally required that humans take responsibility for decisions to use lethal force, and should continue to be in the future.

Distinguishing a target in a field of data is not recognizing a human person as someone with rights. Nor is discriminating between combatants and noncombatants sufficient for recognizing someone as a human being with rights to dignity and life. When it comes to making proportionality decisions, the value of human life is not quantifiable in any deep sense. As human beings who have experienced loss and are ourselves mortal, we have access to the qualitative value of human life. And finally, machines are not capable of deciding questions of military necessity: whether a state of war exists, whether a person in a given situational context can be justifiably killed, or what reasons justify the necessity of destroying a military objective.

There are two types of objections to this framing of autonomous weapons and human dignity. One objection is that machines might be programmed to

perform better than humans in some sense. On the one hand they might be better at discrimination, proportionality, or even military necessity calculations. However, this would depend on a consequentialist analysis, whereby "better performance" consists of making automated choices more accurately or reliably than a human. But this overlooks the reasons and justifications for the decisions, and respect for the underlying duties and rights, as opposed to the consequences, of choices. When it comes to human dignity, what is crucial is both the manner in which the decision is made and the legitimacy of who is making the decision, not simply the final outcome of the decision. The second objection is some form of skepticism of human dignity—either that it does not really exist or that it is reducible to the formal respect of other rights (e.g., life) or that it is a spiritual, mystical, or ephemeral quality that can never be adequately respected and thus can be ignored. While it is notoriously difficult to argue against skepticism, the fact that human dignity has been articulated in various world philosophies and religions, has been integral to legal theory and practice, and has been codified in the constitutions of countries, including Article 1 of the German Constitution and the Preamble of the United Nations Universal Declaration of Human Rights, provides strong support that human dignity has both a defined structure and is broadly recognized as integral to law and morality.

7.3.3. Meaningful Human Control

Given the moral obligation to ensure that weapons are used only against justified and lawful targets, what is the best way to realize this obligation? The discussions by states at the United Nations Convention on Conventional Weapons has repeatedly asserted, with consensus, a few key points in this regard. First, that international humanitarian law applies in all cases of armed conflict and to all weapon systems. Accordingly, the Geneva Conventions, which require commanders to take all precautions to protect civilians in every attack, as well as to review all "new weapons, means and methods of warfare" for their compliance with the law under Article 36 of Additional Protocol 1 of the Geneva Conventions, also apply to any autonomous or highly automated weapons. But how to ensure these obligations are met actually requires that there is space for human reasoning and moral consideration within decisions to use violent force.

It has been proposed at the UN's Convention on Certain Conventional Weapons discussions that what is needed for this is a requirement for "meaningful human control" over the targeting and engagement of all weapons. This could be viewed as the positive obligation that mirrors the negative obligation of not delegating the authority or responsibility for lethal decisions to machines or automated processes. But this term serves other functions and deserves a bit

of unpacking. While "control" does most of the work in terms of moral responsibility, the "human" element is clearly the one that stands out as a requirement for the nondelegation of certain authorities. There is a sense in which software, and automated systems in general, are authored and created by humans and could be seen as a form of human control. While this is acceptable in certain situations, such as automatic doors and thermostats, the decision to use violent force and take human life requires a human capable of assessing the situation, determining the necessity to engage a weapon on a target, who has access to the moral and legal justification for the use of violent force, and who can take moral and legal responsibility for the consequences of that decision. Together these elements add up to something we can describe as "meaningful."

The "control" element implies that systems are acting at the direction and under the control of some specific and identifiable person. This is important both in terms of accountability for the consequences of the use of such a system, but also that there is a human who can actively intervene on a system should it act erratically or unpredictably or create undesirable effects. This is control both in the engineering sense and in the legal and moral sense of there being a human who is controlling a system to achieve the intended results. If the system can act without human intention, or against those intentions, then it is not really in control or is completely out of control.

The "human" element of the term is meant to carry the full burden and responsibilities of moral and legal agency. By virtue of being a morally responsible human, one has an understanding of the value of human life and human dignity that cannot really be represented in a calculation. It is the human condition, and our particular relationship with mortality and life, that informs and grounds our morality. It is the human, or group of humans, who controls a system who is responsible for it and the consequences of the actions taken by the system. Responsibility here operates in two directions. On the post hoc side of things, responsibility means that we can hold somebody responsible for what an autonomous system does, after the fact. For this to be fair, the responsible person must actually be in control of the system. If a system goes haywire and acts in unpredicted and unintended ways, this would tend to diminish the responsibility of the human. Indeed it would be unfair to hold someone responsible for the complex actions of a system that the individual could not possibly have foreseen. But we can hold them responsible if they should have known, or if simply activating a system was itself reckless or negligent.

Responsibility also applies *before* a system is deployed or engaged, in terms of acting psychologically on the human who is in control. That is, one crucial way that moral norms function to regulate human behavior is to encourage people to reflect upon their actions before taking them and to avoid immoral or reckless acts. In psychological terms, the person should consider the moral implications

and consequences of their action and take moral responsibility for choosing to act in certain ways. A key moral issue with automating all sorts of decisions is that it removes human conscience and moral deliberation from the decision process. If an act is morally questionable or morally wrong, it should be hard to commit.

This comes to the "meaningful" portion of the concept. Human acts are social and cultural acts that have meaning within social and cultural systems. War itself is a cultural phenomenon, imbued with complex layers of meaning. For war to be meaningful at all, humans must be capable of engaging with it in meaningful ways. While we do not wish to encourage warfare or to glorify it, we also should not permit violent acts of armed conflict that are untethered to human understanding and morality. Killing should be done only with good reasons, and humans should always be deliberating whether their reasons are good. And when they are, there should be a human willing to take responsibility for acting on those reasons.

It could be argued that we might be able to build artificial agents who are in fact capable of moral agency. If that were possible, would it be acceptable for such entities to make decisions and take actions to kill humans? I am skeptical that we understand the nature of moral agency sufficiently to automate it. While we can create models of moral decision making and teach machines to follow them, or even simulate aspects of human psychology, these are not really the same thing as being a moral agent and taking responsibility for one's own actions. While we cannot prove in principle that it is impossible to replicate the capacities in artificial systems, it would be extremely challenging and would itself be immoral.[30] Furthermore, even if we succeed in creating artificial moral agents, we still might not consider them human persons, but perhaps a more alien form of being, deserving of respect perhaps but not necessarily entitled to judge when it is acceptable to take human life. At any rate, for the foreseeable future we should work to ensure that autonomous systems are not given the authority to use lethal or violent force against humans.

Notes

1. One should be careful in the use of such anthropomorphic terms as "goal" and verbs like "seeks." In humans these terms imply a kind of intentionality that, so far, machines are not capable of. This was the subject of much debate in early cybernetics, as well as in functionalist theories of the mind in philosophy. Peter Asaro, "Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses," *Paradigmi: Rivista di critica filosofica* 33, no. 3 (September–December 2015): 83–107.
2. International Committee of the Red Cross, "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons," ICRC

report, September 1, 2016, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.

3. Examples of such systems include the U.S. Patriot missile system, Phalanx, and Close-In-Weapons-Systems, and the Israeli Iron-Dome system.
4. Ronald C. Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant," *AISB Quarterly* 137 (2013).
5. Robert Sparrow, "Robots and Respect: Assessing the Case against Autonomous Weapon Systems." *Ethics & International Affairs* 30, no. 1 (Spring 2016): 93–116.
6. Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant."
7. Peter Asaro, "What Is an 'AI Arms Race' Anyway?," *I/S: A Journal of Law for the Information Society* 15, nos. 1–2 (Spring 2019): 45–64.
8. There is some debate in political science as to whether arms races are the cause or the result of political instability. I am inclined to view such arms races as involving positive feedback loops, wherein small instabilities lead to small arms buildups, which lead to greater instability and great buildups. See Asaro, "What Is an 'AI Arms Race' Anyway?"
9. In purely economic costs, the Cold War is estimated to have cost the United States $5.5 trillion to $8 trillion in inflation-adjusted military expenditures. One can only wonder what else those funds could have been spent on. See "Cold War's Heavy Cost," *New York Times*, May 20, 1999, https://www.nytimes.com/1999/05/20/opinion/l-cold-war-s-heavy-cost-770728.html; "Effects of the Cold War," Wikipedia, accessed February 23, 2020, https://en.wikipedia.org/wiki/Effects_of_the_Cold_War.
10. Daniele Amoroso and Guglielmo Tamburrini, "The Ethical and Legal Case against Autonomy in Weapons Systems," *Global Jurist* 17, no. 3 (January 2017).
11. Unless, of course, they are equipped with meaningful human control and are designed to seek human authorization for the use of force, and hence are not by definition autonomous weapons in that instance.
12. See "2010 Flash Crash," Wikipedia, accessed February 23, 2020, https://en.wikipedia.org/wiki/2010_Flash_Crash.
13. This is most likely how the Iranian government forced down and captured a top-secret US RQ-170 surveillance drone flying over its territory in 2011. See "Iran-US RQ-170 Incident," Wikipedia, accessed February 23, 2020, https://en.wikipedia.org/wiki/Iran%E2%80%93U.S._RQ-170_incident.
14. Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems," *Journal of Military Ethics* 9, no. 4 (2010); Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant."
15. Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant."
16. Arkin offers only anecdotal evidence and citations to civilian casualties at US checkpoints in Iraq to support this claim. However, we lack any substantial statistics as to how many casualties are due to such "errors" as opposed to other causes. It is also unclear how many civilian casualties are the result of what soldiers or officers consider acceptable proportionality considerations or acceptable levels of collateral damage. It is also unclear how many civilian deaths in conflict zones result from nonweapons, such as loss of clean water, disease, starvation, and dislocation—the plight of those civilians would not necessarily be improved by automated targeting. And, of course,

since the technologies in question are still hypothetical, it is impossible to evaluate their actual performance or compare it to human performance. Moreover they are likely to have their own types of errors and mistakes.

17. Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, *An Ethical Governor for Constraining Lethal Action in an Autonomous System,* Georgia Tech Technical Report GIT-GVU-09-02 (Atlanta: Georgia Institute of Technology, 2009).
18. There are, of course, good reasons to believe that the rules of law are not easily encoded as computer programs. See (Asaro 2009).
19. Peter Asaro, "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics," in *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. Patrick Lin, Keith Abney, and George Bekey (Cambridge, MA: MIT Press, 2011), 169–86.
20. Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007).
21. Peter Asaro, "The Liability Problem for Autonomous Artificial Agents," paper presented at AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University, Stanford, CA, March 21–23, 2016.
22. Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons," *University of Pennsylvania Law Review* 164, no. 1347 (2016).
23. For example, collateralized debt obligations were a means of packaging high-risk subprime mortgages in ways that obscured their risk and diminished or eliminated the responsibility of both the issuers of the original debts and those who repackaged them by appealing to third-party accreditations. The bubble created by these instruments eventually burst, causing the global economic crisis of 2008, for which nearly no individuals or institutions were held liable. See "Financial Crisis of 2007–08," Wikipedia, accessed February 23, 2020, https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008.
24. Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making," *International Review of the Red Cross* 94, no. 886 (Summer 2012): 687–709.
25. Consider the case of an adult leaving a small child in charge of a fire or to operate dangerous machinery. Here we would call the adult irresponsible and immoral, not the small child who is not capable of taking on such responsibilities.
26. Christof Heyns, "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns," A/HRC/23/47, 2013, https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23–47_en.pdf.
27. Christof Heyns, "Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective," *South African Journal on Human Rights* 33, no. 1 (2017): 46–71; International Committee of the Red Cross, "Towards Limits on Autonomy in Weapon Systems," ICRC Statement, April 9, 2018, https://www.icrc.org/en/document/towards-limits-autonomous-weapons.
28. Heyns, "Autonomous Weapons in Armed Conflict and the Right to a Dignified Life."
29. Asaro, "On Banning Autonomous Weapon Systems."
30. Joanna J. Bryson, "Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics," *Ethics and Information Technology* 20 (2018): 15–26.

References

Amoroso, Daniele, and Guglielmo Tamburrini. "The Ethical and Legal Case against Autonomy in Weapons Systems." *Global Jurist* 17, no. 3 (January 2017).

Arkin, Ronald C. "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9, no. 4 (2010).

Arkin, Ronald C. "Lethal Autonomous Systems and the Plight of the Non-Combatant." *AISB Quarterly* 137 (2013).

Arkin, Ronald C., Patrick Ulam, and Brittany Duncan. *An Ethical Governor for Constraining Lethal Action in an Autonomous System*. Georgia Tech Technical Report GIT-GVU-09-02. Atlanta: Georgia Institute of Technology, 2009.

Asaro, Peter. "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George Bekey, 169–86. Cambridge, MA: MIT Press, 2011.

Asaro, Peter. "The Liability Problem for Autonomous Artificial Agents." Paper presented at AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents. Stanford University, Stanford, CA, March 21–23, 2016.

Asaro, Peter. "Modeling the Moral User: Designing Ethical Interfaces for Tele-Operation." *IEEE Technology & Society* 28, no. 1 (2009): 20–24.

Asaro, Peter. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making." *International Review of the Red Cross* 94, no. 886 (Summer 2012): 687–709.

Asaro, Peter. "Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses." *Paradigmi: Rivista di critica filosofica* 33, no. 3 (September–December 2015): 83–107.

Asaro, Peter. "What Is an 'AI Arms Race' Anyway?" *I/S: A Journal of Law for the Information Society* 15, nos. 1–2 (Spring 2019): 45–64.

Bryson, Joanna J. "Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20 (2018): 15–26.

"Cold War's Heavy Cost." *New York Times*, May 20, 1999. https://www.nytimes.com/1999/05/20/opinion/l-cold-war-s-heavy-cost-770728.html.

Crootof, Rebecca. "War Torts: Accountability for Autonomous Weapons." *University of Pennsylvania Law Review* 164, no. 1347 (2016).

"Effects of the Cold War." Wikipedia. Accessed February 23, 2020. https://en.wikipedia.org/wiki/Effects_of_the_Cold_War.

"Financial Crisis of 2007–08." Wikipedia. Accessed February 23, 2020. https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008.

Heyns, Christof. "Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective." *South African Journal on Human Rights* 33, no. 1 (2017): 46–71.

Heyns, Christof. "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns." A/HRC/23/47. 2013. https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23–47_en.pdf.

International Committee of the Red Cross. "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons." ICRC Report, September 1, 2016. https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.

International Committee of the Red Cross. "Towards Limits on Autonomy in Weapon Systems." ICRC Statement, April 9, 2018. https://www.icrc.org/en/document/towards-limits-autonomous-weapons.

"Iran-US RQ-170 Incident." Wikipedia. Accessed February 23, 2020. https://en.wikipedia.org/wiki/Iran%E2%80%93U.S._RQ-170_incident.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 5 no. 3 (2004).

Roff, Heather M. "Killing in War: Responsibility, Liability, and Lethal Autonomous Robots." In Fritz Allhoff, Nicholas G. Evans, and Adam Henschke, eds., *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*, edited by Fritz Allhoff, Nicholas G. Evans, and Adam Henschke. Milton Park, Oxon: Routledge, 2013.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007).

Sparrow, Robert. "Robots and Respect: Assessing the Case against Autonomous Weapon Systems." *Ethics & International Affairs* 30, no. 1 (Spring 2016): 93–116.

"2010 Flash Crash." Wikipedia. Accessed February 23, 2020. https://en.wikipedia.org/wiki/2010_Flash_Crash.

8

Near-Term Artificial Intelligence and the Ethical Matrix

Cathy O'Neil and Hanna Gunn

There are several strands of recent work on AI, including a focus on more abstract philosophical problems, among others: Could AI have genuine emotions? Will the singularity be the end of the species? If we can, should we upload our minds? But there is very important research to be done on person-affecting problems raised by the use of AI systems both in the present day and in the near future. In particular, there is a pressing need to recognize and evaluate the ways that structural racism, sexism, classism, and ableism may be embedded in and amplified by these systems. More generally, there are concerns that the adoption of AI ignores the interests and needs of anyone who isn't part of the development or deployment team.

In this paper we take up the issue of near-term artificial intelligence (AI). "Near-term AI" is used to denote artificial intelligence algorithms that are already in place in a variety of public and private sectors, guiding decisions that pertain to advertising, credit ratings, and sentencing in the justice system. Our focus here is to contribute to a critical discussion of the ways that AI is already being widely used in decision-making procedures in these areas. We will argue that developers and deployers of AI systems—in senses to be defined—bear a particular kind of responsibility for the moral consequences of near-term AI. We will present a tool to aid developers and deployers in engaging in the moral reflection we argue is required of them, in order both to help them to meet their moral obligations and to help address the material risks posed by what we take to be the status quo of actual near-term AI development. This chapter can be understood as a contribution to the field of technology assessment, but instead of suggesting policy revisions, we will propose a framework for ethical analysis that can be used to facilitate more robust ethical reflection in AI development and implementation.

We begin in section 8.1 by introducing near-term AI as algorithms designed as expert systems to replace human decision-makers. This is despite many algorithms being designed as complementary to human decision-makers rather than replacements for them. We then proceed to argue that the current status

Cathy O'Neil and Hanna Gunn, *Near-Term Artificial Intelligence and the Ethical Matrix* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0009

quo in designing and implementing near-term AI doesn't meet minimum acceptable ethical standards because the designers of these algorithms fail to consider the interests of a wide enough range of stakeholders—most significantly, those who will actually be evaluated by these AI systems. We will argue that the present norm that establishes who counts as a designer of an AI algorithm is such that typically only the developers (e.g., data scientists or programmers) and the deployers of the algorithm (e.g., a court, a local government) count. We take this to be problematic, as we will argue it is a primary cause of why the interests of wider stakeholders do not make it into the development of the algorithm, for example, the interests of those who are evaluated or judged by the algorithm. In section 8.2 we argue that we need to develop a wider definition of "success" for near-term AI that better reflects the interests of a wider range of stakeholders. In section 8.3 we discuss a case study on the choice to optimize an AI to different definitions of "fairness"; we show how this decision cannot be separated from ethical decision-making, supporting our argument that designers have moral obligations in the development of AI. In section 8.4 we introduce the ethical matrix framework as a tool for intentionally analyzing the ethical consequences of a new technology. The ethical matrix was proposed by Ben Mepham[1] as a guide for analytic ethical reflection by nonethicists; it typically consists of a 3x4 matrix of three ethical concepts (autonomy, well-being, justice) and four stakeholders. To complete a matrix, one considers how each stakeholder will predictably be affected by the new technology with respect to their interests as represented by the ethical concepts. In section 8.5 we present an ethical matrix that incorporates the language of data science and apply this to a case study. We conclude in section 8.6 with a modified version of the ethical matrix to propose a tool that data scientists can build themselves.

8.1. Problems of Near-Term AI

When we hear "artificial intelligence" we typically think of robots and machines capable of thinking and acting like humans, or, alternatively, of robots and machines that are *far more* intelligent than humans. The thought often continues along dystopian lines, so that these superintelligent machines pose a great threat to humans in one way or another. We will call these "futurist" AI systems, with corresponding "futurist" concerns. At the present moment, though, we do not have superintelligent robots plotting against us. This doesn't mean, however, that there is no artificial intelligence around—the problem is that we're not good at recognizing it. Recent public scandals on the data-trawling business models of social media companies, however, have started to redirect some attention to the AI already in play in many products we use and services we rely on.

Futurist concerns of the sort above will likely resonate with persons most familiar with AI from predominantly science fiction, though of course much serious academic work has also discussed the plausibility and risk posed to humans by AI of the future.[2] While there are numerous academic and researched-based initiatives in place to address a range of issues around the AI presently in use, the status quo in *industry* is still not to engage with the ethical concerns that are becoming more widely recognized by academics and researchers. Some examples of these initiatives are the Campaign to Stop Killer Robots (a conglomerate of NGOs, including Human Rights Watch and Amnesty International), the AI Now Institute, the ACM FAT* annual conferences, and numerous AI labs at universities internationally. Nonetheless our general claim about the intuitive concerns posed by AI stands, that industry standards have largely not adopted ethical goals or interests within their design briefs and that many people are not aware of or concerned about many of the algorithms already involved in making decisions in our lives—despite the ways that they or people they know are affected by them.

Many AI algorithms automate a task previously performed by human workers with expertise or specific training. An automated algorithm can crunch larger amounts of data very fast to deliver a result and thus can either replace a human worker or speed up someone's work. We will call these presently existing algorithms near-term AI, "near" because the cases we are concerned with are either already at hand or are in the process of design and/or implementation. The general blindness to presently existing artificial intelligence has consequences: if we're not paying close attention to the artificial intelligence around us, we're hardly likely to be making sure that it is designed and implemented in ways that—at minimum—meet widely recognized moral standards and avoid inflicting great harms.

A word on terminology is required. Whether or not the examples we discuss qualify as AI for an expert machine learner is irrelevant; from the perspective of the targets of these scoring algorithms, they are sophisticated and opaque black box systems that make important decisions about people's lives. There exists a cluster of ethical problems that arise with automated algorithms that warrant discussing them as a general type, even if we can make more fine-grained distinctions between the varieties of machine intelligence presently in existence.[3] For instance, if an automated algorithm denies one's family medical coverage without warning (or without a meaningful warning)[4] because one failed to check a box on a digital form, it makes little difference whether it's ELIZA or Deep Blue behind the scenes. In both cases, we can ask questions about the decision to design and implement an automated algorithm with the power to remove a family's medical coverage without warning. These kinds of decisions around the design of the automated algorithm, including the choice of data sets,

are ethically problematic aspects of currently existing automated algorithms—whether those algorithms are complex lookup tables or neural networks.

In this first section, we want to bring attention to a number of these ethically problematic issues around the design of near-term AI that we will argue stem from a common source: a failure to consider the interests of many persons who will be deeply affected by the algorithm.[5] We will use the term "stakeholder" to denote those deeply affected by an algorithm; these may be the producers of the algorithm, the deployers of the algorithm, those scored or otherwise evaluated by the algorithm, or those companies or communities whose lives or professions will be disrupted by the widespread adoption of the algorithm. This is not an exhaustive list, and we take up the issue of recognizing stakeholders throughout this piece. We take it that, of all the stakeholders there are with respect to a particular algorithm, only two groups are typically taken into consideration in the design of an algorithm, these being the developers and the deployers. We turn now to an example of an algorithm in Indiana to bring out some of the ways that just who gets to contribute to design can determine whose interests are taken into account and the harms that can accrue to those excluded from this process (and excluded because they are not actually involved in design or excluded because their interests are not considered by others who are in the position to do so during design).

Consider an algorithm designed and implemented for the Family and Social Services Administration (FSSA) of Indiana in 2006, which aimed to modernize the provision of welfare benefits, food stamps, and public health insurance (Medicaid).[6] The goals for the new system were to reduce fraud, reduce public spending, and reduce the welfare rolls.[7] The new system replaced individual caseworkers with an automated eligibility-determining process that used a website for applications and a (privately run) centralized call center to replace one-on-one meetings with caseworkers.

One important factor in deciding who gets benefits is a basic decision about whether to err on the side of minimizing false positives, in this case people receiving benefits they do not need, or false negatives, taking benefits from people who do need them. Prior to modernization, the false positive and false negative rates in Indiana for the provision of welfare were consistent with US national averages, at 4.4% and 1.5%, respectively—a trend in erring on the side of giving benefits to those who don't need them.[8] The goals of the new system erred on the side of producing more false negatives—denying benefits to people who are in need—and so the algorithm was designed with this in mind. Overall the combined error rates rose between 2006 and 2008 to 19.4%, with the greatest rise in benefits denied to people who needed them, a false negative rate of 12.2%.[9] One of the striking features of the new system was a "one-size-fits-all" denial notice: "Failure to cooperate in establishing

eligibility." The language itself, while vague, is strikingly confrontational and accusatory, in effect telling applicants that they have been denied benefits because they are being uncooperative by shirking the rules and regulations of the state. Whether the fault in fact lay with the applicant was not a condition for receiving this notice, and it was not accompanied with any further explanation for why an individual's application was denied. The new system did not make use of existing personal records from the previous system, instead requiring all users to resubmit all of their paperwork. This led to very high rates of lost documentation and a denial of benefits for allegedly "failing to comply."

While it is hard to say that there is a single, central fault in the design and implementation of the new FSSA system, we think it is clear that one of the major downfalls of its design and implementation was a failure to properly engage with the impact of the proposed algorithm on a sufficient range of stakeholders. The goals are explicitly those of the local state: save costs, minimize free riders and fraudsters, and reduce welfare rolls. The new system as a whole is geared toward meeting these goals: reducing staff (by utilizing a private, centralized call center and a computer-based checklist), automating applications (via an online application system), and automatically denying benefits to anyone who makes a mistake or misses a deadline.

Whose interests are not taken into account in this design? Most significantly, the new system isn't concerned with prioritizing—or even affording minimal consideration to—the interests of applicants or caseworkers. The benefits that would seem to accrue with the system are to the state (which may save money), the companies that produced the algorithm, and perhaps to the politicians who fulfill campaign pledges. The developers and deployers failed to determine the predictable ethical consequences of their decision to prioritize the reduction of false negatives for those who would be scored by the algorithm. The result? First, the system denies to people who need public assistance the ability to meet their basic human needs and have access to food, healthcare, and money.[10] The new system also provides little to no person-to-person contact, instead requiring applicants to use the online system, a serious problem for blind or deaf persons who rely on public assistance.[11] Second, the system failed to provide any meaningful level of transparency either to applicants or to caseworkers about how decisions were made regarding the distribution of benefits. Third, the system had serious problems with its data quality by failing to make use of an available database and instead requiring new applications from all beneficiaries. Jane Gresham, a long-term employee with the FSSA, described the new system as "de-humanizing" to both employees and clients. As someone who had been a caseworker with FSSA for two decades prior to the modernization, Gresham described the new system as "factory" work, given the new task-based format,

which undermined workers' abilities to actually oversee a particular client and their needs.[12]

This is a paradigm instance of an ethically problematic near-term AI algorithm. We think that a central failure lies in the way the new automated system was designed with such a narrow focus on the interests of the state. Had the developers, IBM and ACS, and the state been required to consider the interests and needs of caseworkers and those actually dependent on the system as well as those of the state, it is hard to see how this modernized replacement would have been the result.[13] A wider consideration of other stakeholders' interests would make it far more likely someone would have reasoned through the consequences of denying poor families access to food stamps and sick persons their healthcare because they missed an automated phone call. The sheer failure of the system to recognize the costs to real persons' lives in the interests of economic and timely efficiency was remarkable.

We argue that developers and deployers of near-term AI have a moral obligation to engage in ethical deliberation about the consequences of the algorithms they design and deploy. In particular, they have an obligation to engage in a process of determining the predictable consequences of their design choices from the perspectives of those who will predictably be deeply affected by those choices, and to then make an informed decision about how to balance competing interests and values against one another. These include choices about what the purpose or goal of the algorithm is to be (e.g., minimizing welfare rolls) and choices about how the algorithm will meet those goals (e.g., by optimizing to false positive rates). As we will continue to establish, developers (again, those who actually write the algorithms) are in a unique position of responsibility over the design of the algorithm as they are typically the only ones in a position to understand how the algorithm functions and are responsible for rendering the design goals into the algorithm. We will argue for a minimal standard for meeting this obligation, and that is to actually engage in a structured reflection on the predictable consequences of the algorithm by using an ethical matrix. We seek here to establish that such structured reflections are necessary and possible, and we provide a framework for engaging in them. This framework does not require specialist training; rather it asks individuals to apply their commonsense intuitions about, for example, what is fair, in combination with empirical data about the predictable consequences of the algorithm's design.

One of our concerns with near-term AI is that, because the developers are concerned primarily (or exclusively) with their own interests (as a company) or those who will deploy the algorithm (e.g., the state), near-term AI is at great risk of exacerbating harms to already marginalized groups because the interests of those groups are not a part of the conversation around design. Take the FSSA's new system that lacks provision for disabled persons—plausibly an oversight

that we think could constitute ableist structural discrimination.[14] In addition, the presumption that errors in an application indicate that an individual is attempting to engage in fraud or is free-riding can, in our view, plausibly be interpreted as part of a pattern of classist discrimination, that is, a pattern of negatively stereotyping those who are dependent on welfare. This indicates a general lack of attention to the broader social context in which many near-term AIs come into existence. If our intuitions are on point, it also indicates a lack of attention to the ways that certain groups consistently fail to have their needs recognized and taken seriously by contributing to a pattern of failing to consider the needs of members of these groups. The failure of AI developers and deployers to actually engage in thinking through the consequences of their design choices thus maintains these discriminatory patterns in new ways.

We've drawn a distinction between near-term AI concerns and futurist concerns, but we take it that a failure to address our near-term concerns will make it more likely that a variety of futurist existential threats will materialize and that they will be made worse if current discriminatory trends are not addressed. First, they are made more likely because a continued failure to pay attention to the widespread adoption of AI that has been developed without an attempt to address the harmful consequences of its design choices increases the chance that a pernicious algorithm is implemented somewhere. Second, if we are faced with artificial superintelligences beyond human control that lead to existential threats for some large portion of the population, then intuitively it is made worse by our present structural prejudices leading to, for example, a racially discriminatory extermination scenario.[15] A distinct harm that arises in nonsingularity cases, but is no less an existential threat to many persons, is the failure to address the ways that algorithms—like the ones employed in the FSSA example—further materially undermine the poor and may lead to a future of even starker material inequality and a lack of due process for those groups.

The consequence, then, is that we run the risk of widely adopted automated algorithms in our society that make poor people poorer, fail to help the sick, homeless, or otherwise needy, and so put persons lives in serious jeopardy. Continuing to allow for the largely unchecked adoption of automated solutions to social problems could present seriously dystopian situations in a future where goods are distributed only on the basis of lists designed to meet economic or political goals, with no consideration for the nuance of individual needs. While near-term AIs do not seem to present us with human-extermination scenarios, it is no great stretch of the imagination to see how they can lead to dystopian futures where one's very ability to access healthcare, shelter, and food might be due to an inscrutable score provided by the black box of a near-term AI.

Such a scenario is easily preventable if we adopt a norm (or better yet, a policy) of demanding ethical reflection on the ways that different interest

groups or stakeholders will be affected by the implementation of near-term AI by those who are primarily responsible for the design of such algorithms. We make no claim that engaging in ethical reflection on these systems is easy. Someone will have to draw a line somewhere that will leave some people in need of food stamps and medical insurance without the ability to fulfill these needs (keeping the present systems fixed without radical changes to public provisions of these goods). That being said, we think that there is important ethical work that can be done here by establishing and requiring processes that engage developers (the people who build the algorithm with technology) and deployers (the people who use the algorithm once it's built) in a process of ethical and statistical reflection.

Before moving on, we should clarify our terminology. When we use the term "designers," we want it to include—at a minimum—both the deployers of an algorithm as well as the developers. Ideally we'd include other stakeholders as well, or representatives of stakeholders. As we indicated earlier, we take it that in the ideal situation, those who stand to be deeply affected by the adoption of the algorithm would be included in the design of the algorithm. This could take the form of actually including those persons in conversations during the design process, though we argue here for something less: that developers and deployers are required to engage in a process of empathetic design that considers the needs and interests of these groups, and that they then make decisions about design with this information at hand.

It's far from obvious that this would be the definition for "designers" because currently the standard model for corporation or government agency use of algorithms is that a third-party data vendor sells its "black box services" through a licensing agreement that typically doesn't allow the deployer to see the source code or even understand the code even at a basic level. The problem with that is it's harder to trace mistakes and to assign accountability. Indeed another standard element of the legal setup is an indemnification contract that assigns costs of legal settlements to the vendor, allowing the deployer even more dangerous moral distance from the algorithm they use for decisions like who to hire or fire.

So when we suggest that those deployers, who are often currently being kept in the dark about the algorithmic design, should be considered part of the design process, they're not automatically a part of algorithmic design, so their inclusion at this level is rare. And yet to accomplish an ethical and accountable process, we'd argue, deployers will have to be considered part of the design process. That said, as we noted very briefly earlier, there are certain issues that are beyond the average deployer's understanding, namely, the code implementation, often written in computer languages that take years of training to write and understand. This raises the question: How can the overall design process

involve deployers and developers and yet remain accountable to a common set of ethics?

One way in which developers and deployers are similarly accountable is with respect to the values and goals that the algorithm should try to meet, for example, to reduce welfare rolls. In this way, the design team includes both developers and deployers because they have to come to a mutual understanding of which values are to be embedded in the algorithm, and how conflicting values must be balanced. The goal for an accountable algorithmic process would be to split the "ethical decisions" from the "translation of those decisions into code." In other words, the ethical decisions would be made by the entire design team first. Then the development team's job, from the perspective of accountability, would be to faithfully translate those decisions into mathematically precise code. Ideally they would also place monitors into the system to confirm over time that these decisions continue to hold.

For example, if it was decided that the disparity in false negative rates and false positive rates in the FSSA algorithm was an important indicator of fairness of the modernized welfare system, there should be a way to keep track of both rates and ensure they are within a chosen window of uncertainty and tolerance in disparity. Choosing that metric as a fairness indicator, as well as what exactly would be the "window of uncertainty and tolerance," would be choices decided by the entire design team. If we require that this decision must be made in an informed way, then we both help to uncover areas of potential risk and harm to a wider range of stakeholders and we have a more transparent design process for monitoring accountability once it is implemented.

This discussion exposes a difference as well as a commonality between human-created and human-run decision-making processes and automated decision-making processes. They have this in common: they need to be carefully considered in terms of their impact on all stakeholders. They have this difference: black box algorithms are inscrutable to most parties involved in the implementation of and interaction with the automated decision-maker. Therefore they must be audited, preferably continually, to make sure they are functioning as designed and within limits. The development team is uniquely capable to make that happen, which makes them importantly responsible for the ethical consequences of the algorithm and grounds part of their particular ethical obligations with respect to this process.

So far, we've identified several problems with the design and implementation of near-term AI and identified a starting point for fixing them: requiring an inclusive design team to engage in structured ethical reflection on the proposed algorithm that incorporates the interests of a wider set of stakeholders. In the next section, we begin with how common understandings of "success" in data science mask ethical decisions that are made in the design of algorithms.

8.2. A Better Definition of Success

Too often, conversations about machine learning focus on cutting-edge algorithms like the newest chess or Go algorithms, which pit "man against machine" and impress us with the computer's superior memory and learning speed, albeit in a narrow and limited way. We think that the tendency to focus on these types of algorithms—beyond impressing the lay audience—gives the impression that there's a well-defined concept of "winning" or "success" for all algorithms, and that the computer can be taught to understand this definition.

Of course, in the context of Go or chess, those things are true. But they are the exception, not the rule. In any larger, more complex, social setting, which most algorithms inhabit, there is no one definition of "winning." Any definition is inexact and relies on proxies, and often computers end up optimizing to truly ludicrous if not perverse definitions of success, to the detriment of their human targets.[16]

Historically speaking, this tendency to think of algorithmic "success" as "winning" is inherited from the toy universes of games, in that we inherit the very language we use when we talk about machine-learning algorithms in general. Our understanding of success, then, is often narrow and insufficient for understanding whether an algorithm "works" in the messy reality of human social interactions partly because of this inheritance. That simplification of our language allows us to pretend, or assume, that there is a simple concept of success, and that it's one that computers can be taught, given enough data and enough time.

What is in fact happening behind the scenes, however, is that we've set up the algorithm to refer to and optimize to a definition of success that is constructed by and for the algorithm's designers, and that typically ignores the algorithm's other stakeholders. We might, in fact, say that the only stakeholder is the data scientist and perhaps the company for whom the data scientist works, and the only concern is accuracy, profit, or efficiency, depending on what kind of algorithmic context the data scientist works in.[17]

In general, when we have two metrics, A and B, which are distinct, and we optimize to A, we necessarily do not optimize to B. Indeed when we optimize directly to A we end up optimizing directly away from B with very high probability. The extent to which metric B suffers when metric A is preferred depends on how different they are, how much the algorithm matters, and how much of a feedback loop is produced by the choice of metric A in the first place.[18]

Just to give an example, let's choose A and B to be rather close. When the *US News & World Report* magazine decided to rank colleges, it chose a rather weak set of proxies for "quality," which included the rate at which students who applied were admitted, the rate at which students who were admitted actually accepted,

and the reputation of the college according to other college administrators.[19] Let's set A to be that *US News* definition of a "quality" college, and let's consider B to be the definition of quality that a typical high school senior might care about, which would include costs, educational and professional opportunities and connections, location, and prestige. Enormous effort at enormous cost has been put into gaming the ranking system by college administrators, running the spectrum from cheating to building outsized luxury gyms to attract elite athletes. That cost has translated to higher tuitions. However, the *US News* ranking system doesn't care about cost. In other words, choosing a proxy to college quality that is blind to cost and optimizing to it has meant that any quality proxy that is sensitive to cost will be directly punished.

Here's the problem. When data scientists develop an AI, they choose a metric, A, to optimize to in order to determine whether the AI is successful or not. This is problematic in and of itself, as we noted, because this is likely to be too simplistic and will tend to be a very narrow and specific metric that best suits a narrow range of interests, such as profit or statistical accuracy. Given the earlier argument, then, when a data scientist chooses a self-serving metric, A, and we have multiple distinct groups who will be affected by the choice of this metric and whose interests are best served by metric B, then choosing to optimize to A will neglect the interests of the other stakeholders. Thus we should expect that near-term AI will consistently fail to meet the needs of other stakeholders so long as success is determined by a narrow set of data scientist–serving metrics. And this isn't to say that other stakeholders in a given context want the algorithm to be inefficient, inaccurate, or unprofitable—this could be in the interest of other stakeholders too. Rather, other stakeholders may have other interests that are not being taken into consideration and that may need to be weighed against those interests of the designer or of the implementer (whose preference may shape the proxies a designer chooses).

To briefly demonstrate how incredibly harmful this can be, consider an algorithm that predicts child abuse. This is intended to help people who work at a hotline, and indeed those workers deserve to have a data-driven system that enables them to rely on more than their own judgment. The problem lies in the multitude of stakeholders and their accompanying concerns with the outcomes of the algorithm. In particular, if the algorithm is simply optimized for efficiency or accuracy, neither of those will take into account the dire consequences for children or for parents who are inaccurately understood by the algorithm.

We need to expand our understanding of what it means for a given algorithm to work well, and it starts with understanding what "success" means for all the stakeholders with respect to a specific algorithm, not simply those in control of the code. In the next section, we discuss a study that shows some ways that data scientists are working on mitigating racial unfairness in their algorithms. Our

discussion of this study helps to highlight how a data scientist's preferred definition of fairness may conflict with commonsense understandings of fairness. We will not come down in favor of a specific definition of fairness as a consequence of this discussion. Instead we will use our discussion to further motivate the need for a new norm that forces us to intentionally engage in ethical discussions to reveal whose interests are being taken into consideration.

8.3. FICO Scores, Profit, and "Fairness"

Of course, it's not merely deciding what will make an AI "successful" that runs us into ethically consequential choices. There are considerable ethical implications for how we decide to achieve "fairness" in our algorithms too. Moritz Hardt, along with Eric Price and Nathan Srebro presented a case study on how fairness measures can be used with FICO credit scores to determine who gets loans.[20]

Here's the messy real-world context. We know that wealth inequality is correlated with racial groups, and we also know that this is very plausibly because of a long history of racial oppression and structurally racist policies.[21] Thus there is an intuitive sense in which the present wealth distribution is unfair: there are many people with wealth because of systematic patterns of race-based privilege and oppression, and there are many people who are without wealth because of systematic patterns of race-based privilege and oppression. We will use "intuitive fairness/unfairness" in this section to refer to the inequality that is present in these cases of race-based privilege and oppression.[22] Historically, FICO scores were introduced to equalize chances for loans among men and women and among races, which was a way of giving loans to people even if they had fewer economic opportunities. Thus FICO scores were introduced, in a sense, to make economic opportunity more fair.

As Hardt et al.[23] note, FICO scores "are complicated proprietary classifiers based on features, like number of bank accounts kept, that could interact with culture—and hence race—in unfair ways." That is to say, FICO scores themselves are partly the result of systematic patterns of race-based privilege and oppression that can unfairly, in our intuitive sense, evaluate some deserving people as undeserving of loans and vice versa. So while they are an attempt to provide a quantifiable measure of someone's deservingness of a loan, they are highly likely to be subject to human bias and to structural problems of racial oppression.

The Hardt et al. case study investigated five scoring methods, which were set up as follows: if a credit company was given FICO scores and the racial information of would-be borrowers, how might it ensure that its business practices are "racially fair" using that information? Hardt et al. restricted themselves to building a decision engine that would look at someone's FICO score and race

and, depending on whether the FICO score was above or below some threshold, the individual would get the loan or not. Each of the five scoring methods had a different constraint for setting this threshold, establishing four technical definitions of fairness.[24]

At one extreme was "Maximum Profit" with no fairness constraint, whereby a loan-affording threshold was chosen to simply optimize the profit gained from each racial group. Without providing the details, maximum profit was gained when a FICO score threshold is chosen for each racial group such that 82% of that group do not default on their loans. At the other extreme was "Equalized Odds," which requires that the loan-affording threshold be set by determining both the fraction of nondefaulters that qualify for loans and the fraction of defaulters that qualify for loans to be constant across racial groups. Hardt et al. kept track of what the profit margins would look like, how the thresholds for loans would change, and what proportions of the populations by race would end up with loans for each definition of fairness. Importantly, they also considered the incentives that a given definition of fairness would give to the credit company itself.

A third option was the "Race Blind" condition. We would like briefly to contrast the Race Blind option with the Maximum Profit option as they have striking consequences for our intuitive sense of fairness. In particular, the contrast of the two demonstrates how an attempt to be racially neutral can actually have unfair (i.e., racially discriminatory) consequences. The "Race Blind" fairness threshold is set by ignoring racial categories altogether and setting a single loan-affording threshold at which 82% of the whole population will not default.[25] There is a long (and controversial) history of arguments to support the idea that the best way to avoid racial discrimination is to ignore race. This Race Blind loan scoring algorithm is premised on arguments of this sort.

When one applies the Race Blind scoring system, one ends up with a very large population with a correspondingly broad range of FICO scores. As a result, one needs to have a fairly high FICO score to be rated by the scoring system as someone who should get a loan. In a society where racial minorities are less wealthy, the Race Blind system will make it harder for members of racial minorities to get loans because they will, on average, have lower FICO scores. This is the case even though many members of these racial groups with low FICO scores are predictably unlikely to default on their loans. Hence, although this scoring system ignores race, it disproportionately benefits members of some races and disproportionately disadvantages other.

By contrast, if the Maximum Profit scoring system is in place, a threshold is chosen for each racial group such that 82% of the members of that group will not default. Thus members of less wealthy racial minorities, with lower average FICO scores, will have a higher chance to get a loan when the company tries to

maximize profit from them. This is because the algorithm is now more sensitive to the defaulting rates of each racial group. This is striking: a self-serving motive is more fair than one that intentionally attempts to be nondiscriminatory. This is not to argue that the fairest route will always be the one that maximizes profit, but instead it shows that if we do not take the time to seriously consider how our technical definitions of "fairness" (or lack thereof) will impact different stakeholder groups, we can unwittingly bring about unintuitive ethical consequences.

What Hardt et al. considered in their paper only skims the surface of ethical questions about the morality of FICO scoring. Hardt et al. do not discuss whether FICO scores themselves are racist (and we've given reason to think they very well may be) or whether fairness can be determined solely by whether someone would have paid back a loan. For example, perhaps we should instead investigate whether, by giving individual people in a certain subgroup loans even when they might not be able to repay them, the economic status of the whole subgroup goes up because of the added opportunities.

Of course, a larger question has been left unaddressed:[26] When can a given technical definition of fairness, which would require companies to sometimes give loans to people they know might not pay them back but is "good for the group," outweigh the problem of lost profit?

Our "morals of the story" from these discussions about defining "success" and "fairness" when designing a near-term AI system show that there's a strong sense in which we can't actually make these design decisions *without also making moral decisions that impact other stakeholders*. Put differently, developers and deployers are making moral decisions already, but they aren't necessarily identifying them as such and are instead treating them merely as engineering decisions about an algorithm. The status quo in present near-term AI design and implementation is frequently to consider only the very narrow interests of near-term AI developers and deployers, with the result of effectively ignoring the interests of other stakeholders. We've argued that developers, in particular, are in an important position with respect to the accountability for the design and consequences of an algorithm because they are often the ones who are capable of understanding or accessing the system. In addition, they are responsible for translating the goals of the wider decision-making process it will be a part of, for example, giving out healthcare benefits, into the algorithm itself in such a way that it is faithful to the design. We've argued that both developers and deployers ought to be required to engage in a process of ethical reflection that allows them to make informed choices with respect to all of the stakeholders who will be deeply affected by the algorithm. This adds an additional layer of accountability to the design process by making decisions around AI design transparent by essentially requiring a conversation about how to weigh relevant stakeholder interests against one another. Through our discussion of several examples, we've shown that the other

relevant stakeholders will typically include at least those individuals who will actually be judged or scored by the algorithm; later we discuss the selection of stakeholders further.

We shall now introduce and motivate a tool that can be used by design teams to intentionally engage in ethical reflection on their design decisions, and thus become informed about how their choices may predictably impact other stakeholders. Our tool aims to provide a framework for ethical reflection that doesn't require an education in philosophy (not that there'd be anything wrong with that!). The tool is a version of the "ethical matrix" proposed by Ben Mepham[27] that we supplement with the common concerns of data scientists to make it easily usable for those familiar with working in an algorithmic space.

8.4. The Mepham Ethical Matrix

The ethical matrix allows us to engage in ethical reflection on a new technology without having to solve deep ethical problems first, and without specialist ethical training.[28] It does so by requiring us to consider the interests of a range of stakeholders with reference to three general types of moral goods: well-being, autonomy, and justice. In this section, we briefly introduce the motivations for and process of using the ethical matrix, and then discuss the stakeholders and three ethical principles in more detail.[29]

Mepham's initial proposal is heavily influenced by Rawls's early work on how one might adjudicate between the competing interests of persons, which eventually led Rawls to his proposal of the "veil of ignorance"[30] procedure. The three ethical principles that Mepham settles on combine the Rawlsian proposal and the widely used principles of biomedical ethics: autonomy, beneficence, nonmaleficence, and justice.[31] Filling out an ethical matrix requires a process similar to Rawls's famous thought experiment: we are asked to imagine ourselves as each of the stakeholders on our matrix and to consider how we might be impacted by the new technology with reference to the ethical concepts represented in the columns of the matrix. This is a task that takes place before the implementation of the new technology, and so the cells are completed by drawing on the best evidence available for how a given stakeholder may be benefited or be put at risk by the new technology. Thus the task engages us in thinking counterfactually about what is likely to happen if we do or do not adopt the new technology with reference to the present state of affairs. In Table 8.1 we have reproduced a matrix provided in Mepham,[32] analyzing the possible outcomes of genetically modified maize that was designed for herbicide, pest, and antibiotic resistance.

So, for example, in the cell for "Producers" and "Well-being" in Table 8.1, the topics of income and quality of life for farmers are raised. It is possible that after

Table 8.1 The Ethical Matrix

Respect for:	**Well-Being**	**Autonomy**	**Justice**
Treated organism	[N/A for maize]	[N/A for maize]	[. . .]
Producers (e.g., farmers)	Adequate income and working conditions	Freedom to adopt or not adopt	Fair treatment in trade and law
Consumers	Availability of safe food; acceptability	Respect for consumer choice (e.g., labeling)	Universal affordability of food
Biota	Protection of the biota	Maintenance of biodiversity	Sustainability of biotic populations

adopting the genetically modified maize, early adopting farmers will be benefited by higher income; it is also possible that farmers may suffer health risks from the use of the herbicides to which the maize is now resistant. Each cell, then, can be used to raise a possible benefit(s) or risk(s), or both, for a given stakeholder.

Who uses the matrix? The ethical matrix is designed so that it can easily be used by nonethicists to engage in an ethical analysis of a new technology. It can be used in a participatory workshop setting, by an individual,[33] by a research team,[34] and so on. It is completed prior to the implementation or development of a new technology in order to try to consider the possible consequences for a chosen set of stakeholders. As we will discuss later, we think it can also productively be used after a conversation about a new technology to try to map the concerns that were (and were not) raised—we will call this a "discussion-led ethical matrix." This can be a helpful tool for future design because it can highlight areas of concern that were not foreseen.

The stakeholders of the matrix are chosen by considering who will be impacted by the adoption of the new technology. In putting a stakeholder on the matrix, one commits to treating that group as a moral patient for the purposes of the matrix, that is, as someone or something deserving of moral consideration and respect. This first step in setting up a matrix, then, can itself be a collaborative task, and there may be disagreements about who (or what) deserves consideration. Mepham, for example, argued that, at least for agricultural applications, "the environment" ought to always be a stakeholder, partly because its interests are typically neglected by people developing new agricultural technologies. In order to limit the number of stakeholders on the matrix, one ought to make sure that each stakeholder presents a unique set of interests or concerns on the matrix with respect to the new technology. We argued earlier that the groups to be

judged, scored, or otherwise had decisions made about them are stakeholders of near-term AI, as well as the developers and deployers.

The three ethical principles of the matrix are chosen as prima facie ethical principles, that is, principles that are good rules of thumb for moral actions (but, as rules of thumb, can be disputed or outweighed by other considerations). The ethical principle "to promote well-being" can be understood as a combination of the principles of beneficence (acting to promote stakeholders' interests) and nonmaleficence (acting so as to avoid causing harm). This principle represents what is commonly regarded as a utilitarian approach to ethics. The well-being column requires us to think about how the interests of each stakeholder will be promoted or undermined by the new technology. The ethical principle of "autonomy" or "dignity" should be understood as respect for the freedom of the stakeholder. This principle represents a traditionally deontological ethics, and the column requires us to consider how the new technology will promote or limit the freedom of others as self-directing beings who should be respected as such ("ends in themselves"). Finally, the principle of "justice" should be understood as a respect for "fairness." This is a Rawlsian concept of justice as fairness, and fair institutions and policies can be understood as those that are not significantly responsive to arbitrary differences between individuals and that do not disadvantage those whom we already recognize as being disadvantaged.[35] Thus this column asks us to think about what is selected as a relevant feature for the algorithm to use as a proxy or a metric, how the consequences of the algorithm's use will be distributed across stakeholders, and whether everyone's interests are being given due weight in design. These principles, much like the flexibility of stakeholder groups, can be tailored more specifically to the particular technology at hand.

Here are some ways we might apply each principle when analyzing a near-term AI system. For well-being, we might ask: How will each stakeholder be benefited by the use of the algorithm (beneficence)? How will each stakeholder be harmed or put at risk by the use of the algorithm (nonmaleficence)? Are there alternative methods or processes that are less risky for each stakeholder to achieve the desired outcome? For autonomy, we might instead ask: Will each stakeholder have a choice to use or be a subject of the use of the algorithm? How will each stakeholder be able to determine how the algorithm is used with respect to themselves or their interests? Can each stakeholder meet informed consent conditions with respect to the use of the algorithm (i.e., understand how it works so that they can meaningfully take responsibility for its use, or for its effect on themselves or others)? What are the costs if they cannot meet a suitable standard of informed consent? Finally, for justice: Does the algorithm unfairly favor the interests of one stakeholder without promoting the interests of others,

or by undermining the interests of others? Do false negatives or positives harm or benefit the interests of one group but not others? How and why?

If we consider the FSSA algorithm's initial design, we could represent it on a 1x3 matrix that puts the only stakeholder, the state government of Indiana, against how well the algorithm met the design goals (fraud reduction, efficiency, affordability), with no explicit consideration of how it might meet our common moral standards. If we had plotted it on even a simple ethical matrix with *one* additional stakeholder, the welfare clients, we might *at least* have had the designers consider Table 8.2.

The matrix does not in the end tell us how to design new technologies or whether to implement them, but it does cater to a serious need for ethical reflection. Importantly, it does this by requiring that designers engage in exactly what we think was missing from the FSSA process of design: a consideration of how this new technology will actually impact a wide range of stakeholders. The matrix is *analytical*; it helps us to understand what is at risk and what the benefits are in adopting a technology so that we can then make an informed decision about the design of the new technology. The robustness of this process of analysis, though, is contingent on the ability of those completing the matrix to charitably and sincerely consider the issue from the perspectives of the relevant stakeholder groups. Hence we see particular value in trying to have multiple matrixes completed by different stakeholders who will be affected by the new technology—alternatively, in completing a series of matrixes as new developments and research come to hand. The ethical matrix, then, is a tool that can be used to ensure that algorithmic design is actually informed design with respect to the interests of a wider range of stakeholders. It also establishes a more transparent and accountable design process by requiring the development of an actual artifact of a conversation about what the relevant harms and risks are to stakeholders, that is, the ethical matrix itself.

Table 8.2 FSSA Simple Ethical Matrix

Respect for:	**Well-Being**	**Autonomy**	**Justice**
Indiana State	Cost; stability of community	Transparency with respect to product	Fair treatment of clients, workers; risk of fraud
Clients	Provision of basic necessities	Transparency with respect to own case; informed care	Fair treatment in law; accessibility

8.5. Applying the Ethical Matrix to Data Science

As we noted, ethical reflection can be complicated and divisive. In addition, it can be a challenge to determine the morally relevant aspects of an action, a law, or a new technology on which one should focus. This applies to the use of the ethical matrix as well: just what should one consider about a near-term AI when reflecting on stakeholder autonomy? In this section, we bring together the language of typical data science with the ethical matrix. This provides us with a laundry list of items for the substructure of the cells—the things that we ought always to try to consider when analyzing a near-term AI with an ethical matrix.

Data scientists already have metrics by which they assess the quality of an algorithm; these can serve as a list of topics to consider for the potential benefits and risks of a near-term AI. Here is a basic list of familiar data science concerns (that we put to work in the case studies in the next section): profit, fairness (which, as we've seen, needs to be specifically defined), false positives and negatives, data quality, proxy quality, efficiency, accuracy efficacy, transparency, and consistency.

Our suggestion, then, is that data scientists (or deployers, or whoever is analyzing an algorithm) do not need to receive other ethical education in order to use the matrix. An introduction to the method and commonsense ethical principles of the columns of the matrix is sufficient. They merely need to consider their familiar concerns from the perspective of a wider range of stakeholders. For example, they should consider what the specific interests of a loan applicant might be when it comes to the data quality that they are using to train their algorithm. The Hardt et al.[36] study demonstrates that this can be important for members of racial minorities because missing data on minorities can mean fewer loans end up being given to members of these groups. This is because credit companies tend to care about optimizing accuracy for the dominant group more than for minority groups. However, when there is less accurate information, credit companies tend to err on the side of caution, which means more false negatives (loans won't be given out even though they would be repaid).

In the following subsection, we provide a case study of the COMPAS recidivism risk model to show how the ethical matrix can be used to analyze a near-term AI system. We make a point of indicating how data science concerns can be mapped onto the ethical framework.

8.5.1. COMPAS Recidivism Risk Model

Recidivism risk models were introduced as a way to help judges and parole boards rely less on their own judgment when deciding how long a sentence a defendant should receive or whether to grant parole to an inmate. Although

actuarial instruments have been in use for decades, the more recent versions are likely to be more mathematically sophisticated. In particular, they are likely to be "black box algorithms" that are opaque both to the targets and the deployers, in this case, the courts. Given that these algorithms are used in such high-stakes circumstances, it's crucial that we think in an expanded way about what it means for a recidivism risk model to "work well." The ethical matrix can shed light on this matter.

In 2016 ProPublica published an audit, including data and source code, of the COMPAS recidivism risk model, a black box recidivism risk tool created by the private company NorthPointe and licensed for use by the court system in Broward County, Florida.[37] They found, among other things, that black male defendants were much more likely than white male defendants to receive high-risk scores, and moreover that the false positive rate was about twice as high for black male defendants as for white male defendants. White male defendants, by contrast, had false negative rates twice as high as black male defendants.

Given that higher risk scores are associated with longer sentences, there's a powerful asymmetry here. In particular, false positives might lead to charges of civil rights violations, whereas false negatives will not. On the other hand, false negatives are a major concern for the court itself and for the judges, who might worry that they are in danger of being accused of not being tough enough on criminals if the criminals end up committing more crime, an issue that is more salient in parole hearings than in sentencing. So here we have a list of interests for courts, the judges, and defendants with respect to false negatives.

What happened next is interesting: the company that built and sold COMPAS to the courts, Northpointe, issued a statement in response to the ProPublica audit, which basically said that they don't define fairness via false positives, and that by their definition of fairness, something they called "predictive parity," which basically means race-blind risk measurement, or "accuracy equity," which is to say similar areas under respective ROC curves, their score was fair.[38] Essentially, Northpointe's response was that their algorithm was fair by their definition of fairness—a definition that is, however, not common to all of the stakeholders. Absent a compelling argument for why Northpointe's understanding of fairness is the one that all stakeholders in the COMPAS algorithm ought to use, this is an unconvincing defense for not taking seriously the concerns that other stakeholders have about false positives.

The first problem we can identify in this case study, then, is that the COMPAS algorithm was not designed by taking seriously the interests of stakeholders who would be deeply affected by the algorithm: the defendants. Rather, it was developed only with a concern for the interests of Northpointe. The second is the problem with their choice of predictive parity as the only viable understanding of

"fairness." As we discussed when introducing the Mepham matrix, there is room for negotiating a specific understanding of "fairness" or "justice" when analyzing a technology. The problem in this case is that Northpointe's choice of predictive parity is one that other stakeholders are highly unlikely to agree with because it predictably fails to allow for serious engagement with their interests. That is, anyone trying to sincerely consider the interests of black and white defendants would not choose this understanding of "fairness."

To illustrate why predicative parity is a poor choice, we can consider the known problem around missing crime data. Typically we consider the records collected by police departments and court systems as indicative of crime: arrests, reported crimes, charges, and convictions. However, there's good reason to think that these data in general, and arrests in particular, are not good proxies for crime because of the influence of structural racism. For example, the racial disparity in marijuana-related arrests nationally is about 4 or 5 to 1, even though blacks and whites smoke pot at around the same rates, by their own admission.[39] If we develop a tool to guide arrest rates based on these police records, we will end up with an algorithm that will disproportionately criminalize blacks. That is to say, if we were optimizing to predictive parity across race, we'd actually be asking for a higher rate of arrests among black criminals than among white criminals, at least for marijuana-related offenses.

Table 8.3 is an ethical matrix that plots our list of data science concerns from the previous section, with a consideration of a broader range of stakeholders in the Northpointe algorithm. We have added in some non–data science risks and benefits in italics. Given known problems in crime data with respect to race, it is a reasonable assumption that different racial groups will have different interests, risks, and benefits and thus should be treated as distinct stakeholders. We note that we have the advantage of being able to have a fairly fine-grained consideration of the different stakeholder groups because this matrix has been completed after the design, implementation, and review of the consequences of the algorithm. However, given that these racial disparities are known issues in crime data, it is not unreasonable for this particular list of stakeholders to have been selected before the implementation of the algorithm.

The ethical matrix is a tool for analysis. It allows us to map out where various features of our algorithm, for example, how accurate it is, will give rise to different kinds of concerns for stakeholders. But it can be useful to add an evaluative dimension to the matrix too, as a way of highlighting areas of concern in particular. This evaluative dimension can be achieved only once a matrix has been completed, mapping out the predictable areas of risk and benefits to each stakeholder. The evaluative content requires, of course, actually engaging in ethical reflection on each cell to assign it moral weight. Thus the evaluative use of the matrix is a way of recording the results of the deliberative process of weighing

Table 8.3 COMPAS Simple Ethical Matrix

	Well-Being	**Autonomy**	**Justice**
Court	Efficiency, consistency	Transparency; freedom to adopt/ not adopt AI	Efficiency; false negatives; data quality
Black defendants	Maximizes treatment and rehabilitation; minimizes *confinement and punishment*	False positives; transparency	Discriminatory bias, data quality; predictive parity; *respect for civil rights*
White defendants	Maximizes treatment and rehabilitation; minimizes *confinement and punishment*	Transparency	Discriminatory bias, data quality; respect for civil rights
Public	Stability of community	Transparency	False negatives; false positives; data quality; fairness in law
Northpointe	Creative freedom; economic interests	Protection of intellectual property	Predictive parity as fairness; fairness in trade and law

each item for each stakeholder against another, by discounting some concerns (as unlikely or of low significance) and highlighting or selecting others as items to be addressed (as very likely or as very costly to a stakeholder).

We can, for example, color-code the cells of the ethical matrix to provide this evaluative content. So, for example, we can assign white to mean "Don't worry" or "Benefits the stakeholder," light gray to mean "Don't worry too much," and dark gray to mean "Here's where we should worry first." In a given situation that a dark gray cell might translate to, for example, there's a good change that someone's civil rights will be violated ("justice"), whereas in other situations it might simply represent lost opportunity or accuracy. Color codings are by construction overly simplified and are not intended to replace the full, nuanced, and possibly open-ended conversation that each cell represents, but rather a forced "vote" on the status of the ethical consideration by the group of people who are in the conversation. Different people in different conversations could and would draw ethical matrixes with

different color codings, but the result is a more transparent and informed design process.

Here is how we can read off some of the cells on the matrix using the following key to represent our evaluative colors:

- Indicates respect for the principle (white or light gray).
- Indicates infringement of the principle or a negative impact on the stakeholder (light or dark gray).

For reasons of space, we explain in more detail how a few of the cells of the matrix can be interpreted for three of the stakeholders.

Developer Autonomy (Creative Freedom and Economic Interests)

- Northpointe is not subject to unreasonable or burdensome regulations that make producing their product impossible.
- Northpointe can make an economically viable product, and can patent their product to protect it.

Black Defendants' Autonomy (Transparency and False Positives)

- Persons assessed by COMPAS have little to no transparency about the process, the data quality, or how the data are being used in their cases.
- Black defendants, in particular, are at risk of losing their material freedom and self-determination by being wrongly imprisoned as a consequence of high false positive rates.

Black and White Defendants' Well-Being (Maximizing Treatment, Minimizing Punishment)

- Persons assessed by COMPAS who may be in need of help, particularly white defendants, may not be helped to attain it, and people who are not at risk of reoffending, particularly when they are black, are likely to be scored as risky by the algorithm. This has a sum effect of maximizing punishment for many people who do not need it (causing them harm) and failing to provide support for those who do need it (failing to benefit them).

Public Autonomy (Stability of Community)

- COMPAS fails to maintain a safe and flourishing community by keeping persons unlikely to cause crimes in prison, and releasing those likely to cause further crimes.

As noted, this ethical matrix has the benefit of having been written after the development and implementation of the COMPAS tool. Thus we can be confident about some of the risks and benefits in the cells, and we are passing evaluative judgment on the contents of these cells in accordance with critics of Northpointe. Of course, for a new or developing near-term AI these cells will only be our best guess about the benefits and risks of the AI. It can be incredibly useful to complete an ethical matrix after implementation to map out what has been identified as an area of concern. Empty cells can indicate a failure to engage in a sufficiently deep analysis of the consequences of the AI on specific stakeholders.

In the final section, we present a second modified version of the matrix tool to better optimize it for use by data scientists. We show how one can build a discussion-led and evaluative matrix while a conversation unfolds about an algorithm to find areas of specific concern—or areas of contested rich concepts like "fairness." The color coding we introduced here provides this usable and easily digestible method for evaluating the matrix.

8.6. An AI Ethics Tool Data Scientists Can Build Themselves

Given that data scientists and computer scientists are not ethical experts, we think it is reasonable to modify the construction of the matrix in the following ways. In a "data science ethical matrix," we name the columns by the particular metrics familiar to data scientists, and we color-code (for this chapter, on a gray scale) the cells after considering the issue from the perspective of the relevant stakeholder with respect to Mepham's commonsense ethical frameworks. Table 8.4 presents the schematic, although in general, as we will see, there will be more columns.

The point here is that data scientists are more comfortable thinking through the ethics of their familiar metrics than understanding their metrics through the lens of ethical concepts, for instance, thinking through "efficiency" as it pertains to a particular stakeholder, rather than thinking through "justice"

Table 8.4 Example Data Science Ethical Matrix

	Efficiency	**Profit**	**Accuracy**
Stakeholder 1			
Stakeholder 2			
Stakeholder 3			

Table 8.5

	Efficiency	False Positives	False Negatives
Court			
Black Defendants			
White Defendants			

with respect to how the algorithm will affect a stakeholder. When we begin with the familiar data science concepts and require a consideration of how a stakeholder will predictably be affected by design decisions with respect to the data science concept, we start the ethical matrix process on more familiar conceptual terrain. While these amount to the same thing in the end, that is, thinking through the ethics of data science metrics requires using one's ethical concepts, framing the task in the language of data science makes the task more manageable for data scientists.

This flexible framework leads to good news as well as bad news. On the positive side, it allows us to map conversations that have already arisen, such as the conversations outlined in the previous section associated with the COMPAS recidivism model. The ProPublica matrix might look like Table 8.5.

As seen in the matrix, ProPublica's main point was that the black defendants were being unfairly scored higher, as was exposed by the extremely high rate of false negatives. In other words, the corresponding cell is considered at high risk for unethical or unwanted effects and is therefore color-coded dark gray. Similarly, the court is worried about repeat violent offenders being freed through algorithmic error. Since this concern is mitigated by the fact that judges have discretionary power, the cell is colored light gray. On these simplified matrixes that do not have a cell substructure, the choice of light gray will need to be negotiated; we've used it here to indicate "Don't worry too much," but one might use it to indicate "Undecided" or "Too much uncertainty."

The response by Northpointe can similarly be framed by the data science ethical matrix in Table 8.6, coded white because they didn't see any ethical problems, having defined "fairness" differently.

And finally, when data quality is considered, we might have the expanded data science ethical matrix shown in Table 8.7.

Another useful aspect of this construction is that it's easy to locate the trouble spots, whatever one would focus on first. Having said that, it's clear from our discussion that different conversations with different participants would lead to different trouble spots. Even so, it's a useful way of steering a conversation to focus on priorities and ending up with an ethical matrix that

Table 8.6

	Accuracy	Predictive Parity	Accuracy Equity
Court			
Black Defendants			
White Defendants			
Public			
Northpointe			

Table 8.7

	Efficiency	False Positives	False Negatives	Transparency	Predictive Patity	Consistency	Data Quality
Court							
Black Defendants							
White Defendants							
Public							
Northpointe							

makes that conversation transparent and establishes an artifact of accountability of the design decisions.

Now for the drawbacks of this new construction. For example, and probably most important, it's not clear when one has performed a comprehensive job because the simplified matrix doesn't include as much detail. In other words, when does Mepham's original ethical matrix accomplish more than a data science ethical matrix? What are we at risk of leaving out? As a practical tool that will require being presented with a guide, we recommend that users of the matrix ought to begin with Mepham's original list of stakeholders: consumer, producer, the public, and the environment. We should include the environment—think bitcoin applications—and future generations, which would force us to consider long-term feedback loops. Then the design team—ideally consisting at least of developers, deployers, and those scored or evaluated by the algorithm (or a representative for them)—will need to make a decision about what modifications ought to be made to this list.

We might also want to establish a minimum list of metrics to use as columns, including, for example, privacy metrics, transparency, false positives and negatives, and data quality, which was the fatal flaw for the Northpointe discussion. Depending on the context or industry, those lists of concerns would vary. If, for example, due process rights were enforced for recidivism-risk algorithms, transparency would become a required consideration for such algorithms. We also suggest that the Mepham commonsense ethical principles autonomy, well-being, and justice be presented as a guide for more comprehensive and explicit ethical reflection. It's entirely possible that one might think about ethics only as a matter of weighing positive and negative utility, for instance, and leave off rights as an issue of justice with which particular stakeholders might be concerned. Given that one cannot complete the evaluative use of the ethical matrix without making judgments about just whose interests matter and to what degree, requiring an explicit framework of ethical principles that requires a rich reflection on a variety of standard ethical concerns (well-being, self-determination, and rights) is a way to make this process transparent—and of course, more realistic to the interests of the stakeholders.

The simplest explanation of a data science ethical matrix is that it's an artifact of a conversation related to a specific algorithm used in a specific context. For our last example, we will map Virginia Eubanks's case study of the Allegheny Family Screening Tool (AFST) model, which predicts child abuse, taken from chapter 4 of *Automating Inequality*.[40] This will certainly not contain all the details she provides, but it should give an overview of the issues that are raised in the chapter using a data science ethical matrix.

The AFST algorithm was built and is used by the Allegheny County Office of Children, Youth, and Families (CYF) in Pittsburgh, Pennsylvania, to determine which of the many calls on the child abuse hotline should be followed up with a caseworker. That immediately gives us the following stakeholders: the CYF office, the parents or caregivers of children who may be suspected of abuse, and the children who are at risk of abuse. We may also want to consider the people who made the call concerning suspected abuse, since that group's perception of the system's working or being flawed will probably contribute to its success. We also might want to include the public at large, since the AFST algorithm is intended to protect the safety of children in the community, and we might want to further split these categories, depending on how the conversation proceeds.

Eubanks goes on to point out that the algorithm will act differently on those families that have had higher interaction with the social safety net, such as homeless families and families already in the foster care system. In particular, the more data that are available about a given family, the more likely it is that their score will be sufficiently high to warrant follow-up.[41] That implies we might want to differentiate between "high-touch families" and "low-touch families." It

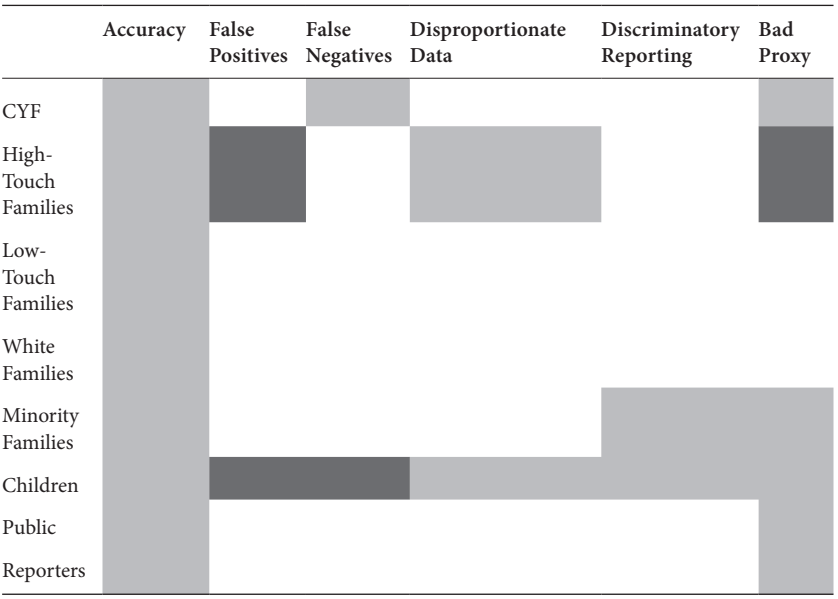
might be easier to simply define "high touch" to mean that a certain threshold of data has been exceeded, or a proxy of low versus high income might make more sense, depending on what information we actually know about the families themselves. The associated concern in the matrix might be characterized as "disproportionate data availability." When we do this, we should keep in mind that one reason there might be disproportionate data availability is that there is a long history of actual abuse, for example. A primary data-driven goal will be to try to distinguish between that "true signal" of abuse and incidental data collection; a secondary data-driven goal, which is famously difficult, would be to try to measure the missing data associated with families that are well-off and that have been historically outside of the system.

Next, Eubanks points out that the way the calls come into the hotline are racialized in general. This suggests that we should further distinguish by race among stakeholders, and that an associated concern would be "discrimination in reporting." We want to make clear here that if the design team had been required to complete an ethical matrix, this would require a process of determining the predictable consequences of the algorithm. In order to do this, one would need to consider the data quality, existing patterns or trends in the population that may be relevant to metric selection, and many other statistical facts relevant to the problem to be solved. That is, proper use of an ethical matrix requires empirical research into what we know about a problem already, and thus is likely to turn up evidence of biased data sets or sampling problems with the target population that the algorithm will be used with. We make these comments to establish that a number of these consequences of the actual design of the AFST algorithm that we can see after implementation may have been predictable had the design team been required to consider them.

Eubanks also makes a salient point about the choice of definition of success for the model; namely, the model doesn't actually train on "substantiated abuse" events but rather either the removal of a child from his or her home or follow-up calls to the hotline. The latter type of event is, once again, known to be racialized, and the former is known to happen to parents who are simply too poor to provide their children with common comforts. In general, according to ambiguous laws, it's difficult to know when to call something neglect and when to simply describe it as poverty. The end result is that we should certainly add a column of concern entitled something along the lines of "target variable imperfect proxy for substantiated abuse," or "bad proxy" for short, and keep in mind that both poor families and minority families are likely to have a bigger problem with this particular column than richer white families.

Altogether we now have a data science matrix that looks something like Table 8.8.

Table 8.8



8.7. Summary

In many cases, the very problems that near-term AIs are being deployed to solve are moral problems: informing decisions about parole and imprisonment, helping to decide who gets loans, determining who is eligible for welfare. Given this fact, it is surprising that ethical reflection is not the norm in algorithm design. In the infamous Trolley thought experiment, an uncontrollable train travels down the tracks on its way to run over and kill five people. You as the observer have the option to divert this train onto an alternative track, where it will instead kill just one person. Some people argue that they can absolve themselves of the situation by refusing to decide whether or not to pull the lever. Perhaps some people developing near-term AIs take this to be their own situation and that their nonengagement in explicit ethical decision-making in AI design absolves them of any responsibility for the consequences that the algorithm has on the lives of the stakeholders assessed, evaluated, or scored by the system. Of course, whether one can really be passive in the Trolley scenario is controversial. We've argued here that design decisions about algorithms are moral decisions; thus we have in a way denied that one can remain on passive or neutral moral ground in AI design and implementation.

As our discussions of a range of case studies show, near-term AI is being implemented in ways that have serious ethical consequences for many persons simply because their interests are not being taken into consideration in the design of these algorithms. We think that this is not only harmful but that it constitutes a widespread failure to meet our ethical obligations not to cause harm to others—a sentiment that is widely regarded as a commonsense ethical platitude. The ethical matrix and the data science ethical matrix are practical tools that can help to fulfill this much-needed space for serious ethical reflection, and we think something of this sort ought to be required of near-term AI design teams. The immediate, and significant, advantage of adopting this tool is that it forces us to consider how our choices affect a range of stakeholders wider than those of the designer and the implementer. Furthermore, it does so in a way that makes AI design processes more transparent to the stakeholders in general, and it assists in making accountability more transparent as well. Even this small step will be a significant improvement for the field of near-term AI.

Notes

1. Ben Mepham, "A Framework for the Ethical Analysis of Novel Foods: The Ethical Matrix," *Journal of Agricultural and Environmental Ethics* 12, no. 2 (2000): 165–76, doi:10.1023/a:1009542714497.
2. On the risks of AI superintelligence and value alignment, see N. Bostrom, "Ethical Issues in Artificial Intelligence," In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, ed. I. Smit et al., 12–17 (International Institute of Advanced Studies in Systems Research and Cybernetics, 2003). On the moral decision-making of autonomous weapons systems, see D. Purves, R. Jenkins, and B. J. Strawser, "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons," *Ethical Theory and Moral Practice* 18, no. 4 (2015): 851–72.
3. We are, in a sense, defining AI by its functional role in our lives—as those automated systems that are typically black boxes to those evaluated by them, and that are included in (or just function as) the decision-making procedure. One could define AI by another metric, for example, the complexity of the processing of the system, whether it uses certain decision-making procedures, whether it is composed of neural networks, and so on. But our concern is with the adoption of algorithmic solutions to decision-making in a wide range of areas.
4. As we will discuss in the coming paragraphs, an automated system adopted in Indiana gave the blanket notice "Failure to cooperate in establishing eligibility" to all persons whose benefits were denied or taken from them.
5. Though it's worth noting, too, that the environment and nonhumans may also have important stakes in the adoption and design of these algorithms. Our scope here will

be to focus primarily on the harms to persons; we do note in later sections where nonpersons might be taken on as stakeholders.

6. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018), ch. 2.
7. A reduction in welfare rolls seems to be serving as a proxy for an improvement in the material well-being of those who had depended on them, and also as a proxy for a more efficient (cost-effective) welfare system.
8. Eubanks, *Automating Inequality*, 48.
9. Ibid., 63.
10. As stated in Article 25 of the Universal Declaration of Human Rights, "Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control."
11. Eubanks, *Automating Inequality*, 68.
12. Ibid., 62.
13. In 2010 Indiana actually sued IBM for a breach of contract for the high rates of benefit denials to citizens in need. The judge found in favor of IBM, finding that the company had met the goals laid out by the design brief: "The heart of the contract remained intact throughout the project" (ibid., 75). As Eubanks describes the situation, and we strongly agree, "The problem . . . was not that the IBM/ACS coalition failed to deliver, it was that the state and its private partners refused to anticipate or address the system's human needs" (75).
14. Overt ableist discrimination would require that someone intentionally designed the system so that it did not cater to the needs of the blind or deaf. We take it that the design failure was an oversight that arose from not considering differently abled persons' needs. For a discussion of "overt" and "institutional" racism that informs this distinction on overt and institutional ableism, see G. Ezorsky. *Racism and Justice: The Case for Affirmative Action* (Ithaca, NY: Cornell University Press, 1996), ch. 1.
15. Our thanks to Michael P. Lynch for raising this point.
16. To be clear, serious academic research outside of this realm of well-defined success, and toward the project of aligning machine values to human values, is underway and exploding in numbers, conferences, and academic output. See, for example, the FAT conferences mentioned earlier, which are dedicated to issues of fairness, accountability, and transparency in machine learning and other automated systems.
17. Academic researchers tend to be given leeway when it comes to how they define success than, say, a data scientist working at a social media company, where the definition of success is likely strictly tied to business interests and held there by corporate lawyers.
18. This even has a name when applied to business metrics: Goodhart's Law. And even though it's widely known, it's still in effect.
19. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2017), ch. 3.

20. Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *arXiv*, October 7, 2016, https://arxiv.org/abs/1610.02413v1. See also O'Neil, *Weapons of Math Destruction*, ch. 8.
21. See, for example, Mehrsa Baradaran, *The Color of Money: Black Banks and the Racial Wealth Gap* (Cambridge, MA: Belknap Press of Harvard University Press, 2017), for a detailed discussion of the history of the racial wealth gap in the United States.
22. The thought being that by commonsense or intuitive uses of "fair play," this is an unfair situation.
23. Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," 17.
24. The first scoring system, "Maximum Profit," explicitly states that it has no fairness constraint. As we will show, however, despite this, it achieves a result with an intuitive sense of fairness better than one of the constraints that seeks to establish a technical definition of fairness as "racially neutral treatment."
25. "We . . . consider the behavior of a lender who makes money on default rates below this, i.e., for whom false positives (giving loans to people that default on any account) is 82/18 as expensive as false negatives (not giving a loan to people that don't default). The lender thus wants to construct a predictor $\hat{y}$ that is optimal with respect to this asymmetric loss." Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," 17.
26. Although to some extent followed up by Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt, "Delayed Impact of Fair Machine Learning," *arXiv*, April 7, 2018, https://arxiv.org/abs/1803.04383.
27. Mepham, "A Framework for the Ethical Analysis of Novel Foods."
28. D. Schroeder and C. Palmer, "Technology Assessment and the 'Ethical Matrix,'" *Poiesis & Praxis* 1, no. 4 (2003): 295–307, doi:10.1007/s10202-003-0027-4.
29. For a more recent presentation of the matrix and a discussion of the ethical principles used, see Ben Mepham, "Ethical Principles and the Ethical Matrix," In *Practical Ethics for Food Professionals*, ed. J. Peter Clark and Christopher Ritson (Wiley Online, June 7, 2013), 39–56, doi:10.1002/9781118506394.ch3. C. Kermisch and C. Depaus, "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices: The Case of Geological Disposal for Radioactive Waste," *Science and Engineering Ethics* 24, no. 1 (2017): 29–48, doi:10.1007/s11948-017-9882-6, assess the utility of the ethical matrix framework, which is typically a collective activity, for individual researchers. K. K. Jensen et al., "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix," *Science and Engineering Ethics* 17, no. 3 (2010): 425–45, doi:10.1007/s11948-010-9218-2, present results that show the ethical matrix successfully works as a tool for promoting ethical reflection among scientists, and in particular on the needs of external stakeholders.
30. (Rawls 2000, 167). In brief and leaving aside a lot of interesting detail, the veil of ignorance is a thought experiment that asks us to imagine that we have been tasked with making political decisions about how to structure and regulate our society. However, we must make these decisions imagining that we have no knowledge of who we are within our society: our social or economic status, interests, religion, level

of education, talents, and so on. The thought is that without selfish reasons to guide us, we are more likely to decide on principles and structures that are truly just or fair.

31. T. Beauchamp and J. F. Childress, *Principles of Biomedical Ethics,* 7th ed. (Oxford University Press, 2012).
32. Mepham, "A Framework for the Ethical Analysis of Novel Foods," 170.
33. See, for example, Kermisch and Depaus, "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices."
34. See, for example, Jensen et al., "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix."
35. Our "intuitive fairness" concept from section 8.2.1 can be understood more precisely with this Rawlsian notion.
36. Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."
37. Julia Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
38. William Dieterich, Christina Mendoza, and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," Northpointe Inc. Research Department, July 8, 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html. The area under ROC (receiver operating characteristic) curves measures the extent to which, as we vary thresholds, the model maintains a better-than-random true positive to false positive ratios. Thresholds are the limit values between the "high-risk" category and the "low-risk" category. Note that, once you are actually using an algorithm, you have of course *chosen one* threshold, so you don't actually care about how accurate the model would have been with differently chosen thresholds, just how accurate your model is with the specific threshold you've chosen. This gives you a bit of a flavor as to why this area is mostly irrelevant to the person deploying the model, never mind the target of the model who is worried that they, specifically, represent a false positive or a false negative.
39. Angwin et al., "Machine Bias."
40. Eubanks, *Automating Inequality*.
41. Ibid., 146, 155.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23, 2016. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baradaran, Mehrsa. (2017). *The Color of Money: Black Banks and the Racial Wealth Gap*. Cambridge, MA: Belknap Press of Harvard University Press, 2017.

Beauchamp, T., and J. F. Childress. J. F. (2012). *Principles of Biomedical Ethics.* 7th ed. Oxford University Press,Newyork 2012.

Bostrom, Nick. (2003). "Ethical Issues in Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2,

edited by I. Smit George Eric Lasker, Wendell Wallach, Iva Smit., 12–17. International Institute of Advanced Studies in Systems Research and Cybernetics,Windsor 2003.

Dieterich, William, Christina Mendoza, and Tim Brennan. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." *Northpointe Inc. Research Department*, July 8, 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Eubanks, Virginia. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press, 2018.

Ezorsky, G. (1996). *Racism and Justice: The Case for Affirmative Action*. *Ithaca, NY*: Cornell University Press, 1996.

Hardt, Moritz, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." *arXiv*, October 7, 2016. https://arxiv.org/abs/1610.02413v1.

Jensen, K. K., E. Forsberg, C. Gamborg, K. Millar, and P. Sandøe. "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix." *Science and Engineering Ethics* 17, no. 3 (2010): 425–45. doi:10.1007/s11948-010-9218-2.

Kermisch, C., and C. Depaus. "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices: The Case of Geological Disposal for Radioactive Waste." *Science and Engineering Ethics* 24, no. 1 (2017): 29–48. doi:10.1007/s11948-017-9882-6.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." *arXiv*, April 7, 2018. https://arxiv.org/abs/1803.04383.

Mepham, Ben. "Ethical Principles and the Ethical Matrix." In *Practical Ethics for Food Professionals*, edited by J. Peter Clark and Christopher Ritson, 39–56. Wiley Online, June 7, 2013. doi:10.1002/9781118506394.ch3.

Mepham, Ben. "A Framework for the Ethical Analysis of Novel Foods: The Ethical Matrix." *Journal of Agricultural and Environmental Ethics* 12, no. 2 (2000): 165–76. doi:10.1023/a:1009542714497.

Mepham, Ben, M. Kaiser, E. Thortensen, S. Tomkins, and K. Millar. "Ethical Matrix Manual: Agricultural and Forest Meteorology." *Agricultural and Forest Meteorology* (January 2006). https://www.researchgate.net/publication/254833030_Ethical_Matrix_Manual.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Broadway Books, 2017.

Purves, D., R. Jenkins, and B. J. Strawser. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18, no. 4 (2015): 851–72.

Schroeder, D., and C. Palmer. "Technology Assessment and the 'Ethical Matrix.'" *Poiesis & Praxis* 1, no. 4 (2003): 295–307. doi:10.1007/s10202-003-0027-4.

9

The Ethics of the Artificial Lover

Kate Devlin

The past few years have seen a cascade of headlines about sex robots. These have ranged from the melancholic ("Seedy, sordid—but mainly just sad")[1] to the outraged ("Sex Robots Could Reveal Your Secret Perversions!")[2] and the wryly alarmist ("Sex Robots May Literally Fuck Us to Death.")[3] By and large, the headlines are dystopian. The overwhelming majority of them assume that a sexual companion robot will take a life-like robot form, usually female. John Danaher[4] proposes a clear definition of sex robot as "any artificial entity that is used for sexual purposes" and that meets three conditions: first, it takes a humanoid form; second, it has human-like movement or behavior; and third, it has some degree of artificial intelligence. These conditions, he notes, can be disputed. He is agnostic as to whether or not such robots should be embodied. Danaher and Neil McArthur's 2017 edited collection *Robot Sex*[5] provides a clear and thorough debate on the social and ethical implications of sex with robots. In this chapter I wish to extend the debate into a new area: the idea of the nonhumanoid sex robot.

9.1. Background

In many of the articles on sex robots, the history tends to begin with the story of Pygmalion. The plot is well-known and one that has been used down the years to popular acclaim, in Shakespeare's *Winter's Tale*, the eighteenth-century opera *Pygmalion*, the ballet *Coppélia,* and the musical and film *My Fair Lady.* The familiar account is that a Greek man, a sculptor called Pygmalion, could find nothing good in women, so he instead created a beautiful statue. He fell in love with the statue and prayed to Aphrodite, the goddess of love, that he would find a woman like the one he had created. Returning home, he kissed the statue and she came to life. They married and had a child.

GenevieveLiveley[6] clarifies that this story originates in the Roman poet Ovid's *Metamorphoses*, dating to 8 CE, and that, as with other stories in the *Metamorphoses*, its theme is deception. It focuses first of all on the delusional character of the main protagonist. He fools himself into thinking a statue made

Kate Devlin, *The Ethics of the Artificial Lover* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0010

of ivory comes to life. Ovid takes great pains to emphasize the fact that this is delusional. Pygmalion's desire was for a real and perfect woman. He wanted what the statue represented.

There are earlier myths dating back to the Greek period that explore the idea of the artificial lover, including the story of Laodamia and Protesilaus (which, interestingly, features a male "robot" as the object of desire) and Polybius's account of a realistic automaton owned by the Spartan king Nabis, a realistic robot designed and dressed to look like his dead wife, Apega.[7] The primal archetype could be said to be Pandora, the first mortal made by the Greek gods, programmed by a team to look and behave in a certain way. This idea of creating artificial humans has a long history.

Our perception of the sex robot as an alluring, seductive, attractive woman is therefore fueled by years of influence from science fiction books and films. In today's popular culture it starts with Maria, the character in Fritz Lang's 1927 film *Metropolis.* Maria, the beautiful heroine, brings hope to the exploited workers with visions of a better future, but her prophecies unnerve the leader of the city and he orders a *Maschinenmensch*—a robot double of Maria—to deceive the proletariat. While she is not a sex robot per se, Maria's beauty and passion are central to the plot: at one point Robot Maria performs as an exotic dancer. It is said to be the first portrayal of a robot in film, and her introduction sows the visual seeds of the widely recognized, beautiful but dangerous fembot—a robot with sexual characteristics.

9.2. The Current Reality

At the time of writing, there are around fifteen workshops worldwide making what they describe as "sex robots," although they are, at best, dolls with limited responses. The closest anyone has come to commercially producing something more technologically advanced is Realbotix, a division of the California-based Abyss Creations, creators of RealDoll.[8] Realbotix have created a version of their silicone doll with an animatronic head and an AI personality. Their robot, which is completely stationary from the neck down and cannot stand unsupported, is called Harmony. The head is interchangeable, and Harmony's personality can also be used as a standalone AI chat app on an Android smartphone or tablet (where Harmony responds in a gentle Scottish accent). Harmony is now available to order, along with a variant known as Solana. At the time of writing, the first completed model is due to be shipped to the customer.

In addition to Realbotix, there are also the "garage builders," like the engineer Sergi Santos. Santos, who is based in Barcelona, began engineering sex robots following a PhD in nanotechnology at the University of Leeds. His robot,

Samantha, is said to be "capable of enjoying sex," with sensors embedded into her body to facilitate specific patterns of touch and the ability to generate emotional states and responsiveness.[9]

Although Realbotix have announced to the press that they will make a male version of their robot,[10] the sexual companion robot is consistently portrayed as female. Gynoids are designed to play to cultural stereotypes, generally taking an eroticized form—shapely, sexy, and obedient. There's an essence of the femme fatale about some of them too—the perfect woman but with the underlying potential for danger. We see the gynoid appear from generation to generation, for example, *The Stepford Wives* in the 1970s, *Weird Science* in the 1980s, *Eve of Destruction* in the 1990s, *Buffy the Vampire Slayer* in the 2000s, and the *Westworld* reboot of 2016. In each, perfect human-like women are built by men as companions and lovers. Wosk gives a comprehensive overview and chronology of the artificial woman.[11] One notable exception is the 2001 Steven Spielberg film, *A.I.*, featuring Gigolo Joe, a male sex worker robot programmed with the ability to mimic love. It should be noted that this rare gesture of sex robot equality did not extend to Gigolo Joe's being portrayed in as sexualized a manner as the majority of fembots.

9.3. Sex

In this chapter I am using the word "sex" as shorthand for "sexual activity." We each carry our own understanding of what the word "sex" means, and there is every chance that we mean different things. The Kinsey Institute for Research in Sex, Gender, and Reproduction carried out a survey in 2010 of more than five hundred people in the state of Indiana, where the Institute is located, to find out what they classified as "had sex."[12] Almost all respondents classified penile-vaginal intercourse as having sex, and 80% said that penile-anal intercourse also counted as sex. Numbers varied significantly, however, depending on age group, when asked about activities such as manual or oral stimulation of a sexual partner's genitals. Masturbation was not included in the survey.

MarkMigotti and Nicole Wyatt provide a dissection of what it means to have sex, incorporating factors such as activities, agency, and the requirement for a partner of partners: the shared agency of "being sexual together" with a partner (or partners) who are subjects rather than objects.[13] Masturbation is not included in their definition, as it lacks the sexual "we." Solo sex, they state, would reduce sex robots to mere sex toys. They argue that if sex robots are simply masturbation aids, then they "don't raise any distinctive social, ethical or conceptual problems." Likewise, Kathleen Richardson writes, "It's only in sexual encounters with others that we can learn the depth of sexual feeling,"[14] although she does not

give a basis for this statement. McArthur describes sex robots as a "special kind of object . . . more than a mere autoerotic act"—contingent on their responsiveness and projection of personhood.[15] But how key is such personhood?

I agree that we may adopt a definition of sex that encompasses many different kinds of acts. It certainly is not limited to penile-vaginal intercourse and does not need to involve penetration or orgasm. However, I disagree that it needs to be a partnered activity. In very basic biological terms, sexual activity could be classed as the act of heterosexual mating, but sexual behavior in today's societies is by no means limited to reproductive purposes (as is acknowledged by Migotti and Wyatt). However, I wish to approach the classification of sex in this instance from a (neuro)biological stance, namely, as any action that causes a feeling of arousal. This, therefore, includes masturbation, which I would class as a legitimate form of sexual activity, which could include both sex toys and sex robots. Indeed I would argue that sex robots *are* a version of sex toys, albeit a more embodied form. They come from the lineage of the sex doll, but they are still only objects, even if they are humanoid. Admittedly, that human-like dimension raises interesting questions about object attachment and rapport with robots—more of which later. However, classing sex robots as sex toys does not preclude social, ethical, or conceptual problems. Those problems still exist, albeit from other angles. It is possible to say that sex robots in their current humanoid form are in some ways distinct from current sex toys—but they also share many similarities.

Again, I concur with Migotti and Wyatt that sexual assault, such as coercive sex and rape, should not be classified as sex. Rape is not consensual. It involves a sexual act, but it is a sexual act of violence: a criminal act of an abuse of power. Rape is often discussed in the debates on sex robots, and it will be discussed in the next section. But for now, I am using the word "sex" to describe a *wanted* act—something that leads to pleasure for those involved.

Modern biological investigations of the sexual response cycle date from the 1960s and the work of William H. Masters and Virginia E. Johnson,[16] who identified a cycle of excitement, a dynamic plateau of surges of pleasure and orgasm and resolution. This led to a number of models of sexual response. John Bancroft defines sexual arousal as "covering a state motivated towards the experience of sexual pleasure and possibly orgasm, and involving (i) information processing of relevant stimuli, (ii) arousal in a general sense, (iii) incentive motivation and (iv) genital response."[17] A more concise definition by James Pfaus is "Physiologic sexual arousal in all animals can be defined as increased autonomic activation that prepares the body for sexual activity."[18] Those sources of arousal could be any combination of primal evolutionary cues, neural responses, and triggers formed by experience and expectation. Desire is distinct from arousal, Pfaus reports, as it is a psychologic interest. While there are certain baseline contributors that

seem involuntary (visual erotic imagery and arousing odor, for example), different minds also respond to different triggers. Frederick Toates' 2009 model suggests an incentive-motivation theory whereby cues, both external and internal, prompt sexual motivation.[19]

In general, it's agreed that chemicals in the brain are responsible for inducing a state of sexual responsiveness. These are the hormones and neurotransmitters that influence our behavior and include oxytocin, dopamine, serotonin, norepinephrine, and melanocortins. These in turn cause physiological responses, such as changes in blood pressure, heart rate, respiration, and genital responses. Functional neuroimaging studies report the brain activity associated with arousal of participants when viewing sexually explicit images, that is, with no human present to trigger desire—as, of course, is the case with anyone viewing pornographic images or video. (Interestingly, in one fMRI study, women showed arousal when shown nonhuman stimuli.[20] Men, however, did not.[21])

Taking arousal and sexual response as a baseline, there seems no reason to exclude masturbation from the category of sexual activity. A human sexual partner (or partners) does not need to be present for sexual excitation to occur: the brain chemistry still responds. Sex robots in a human-like form represent, or mimic, a human-like mutual sexual encounter, but I wish to suggest that this could be extended to nonhuman-like forms of robots, or even to disembodied AI.

9.4. Ethics: Social Harm

Arguments against sexual companion robots hinge on the hypothesis that sex robot–human relationships will be detrimental to society, damaging human-human relationships. Richardson postulates this as a parallel to the sex worker–client relationship, which she views as damaging.[22] Her view of sex work is profoundly negative, arguing from an abolitionist perspective similar to Catharine Mackinnon's[23] and Andrea Dworkin's.[24] Richardson maintains that there can be no consensual participation in prostitution (the term she prefers to use) as it is an act that subordinates (predominantly) women, who have often been forced into the position of selling sex due to adverse life circumstances. Richardson believes that sex robots are modeled on the prostitution dynamic and so, by extension, are similarly problematic. This, she argues, could lead to increased objectification and increased sexual violence and rape. She has called for a ban on the development of sex robots, later modified to a request for "the development of ethical technologies." Her position paper makes explicit reference to trafficking, seemingly conflating this with sex work.

Danaher et al. have provided a thorough analysis of Richardson's arguments, describing her premise as "weak."[25] At a very basic level, her suppositions are

countered by sex-positive feminists who state that describing all sex work as nonconsensual removes agency from women and that, although exploitation and trafficking are clearly wrong and should be eradicated, the sex industry should not be wholly condemned on this basis. Indeed the United Nations Trafficking in Persons Protocol classes trafficking and sex work as distinct phenomena.[26] Between these opposing feminist views is consensus that sex workers should not face criminal sanctions, that authentic consent is a necessity, and that violence and coercion should be eradicated.[27]

DavidLevy, whose 2007 book, *Love and Sex with Robots*, established the study of sexual companion robots as a field of study, has a much more positive view and sees the sex worker–client model as one that can provide benefits, and that those benefits could be mirrored in a sex robot–client relationship.[28] While this may be viewed as somewhat utopian, it has as much going for it as Richardson's argument: the evidence is conflicted, and each side can produce sources to support their view. While I agree with Richardson that perpetuation of the objectification of women is a risk with current forms of prototype sex robots, I find the link to increased sexual violence and rape tenuous. First, we have seen these arguments with computer games: that violence in computer games will lead to an increase in real-life violence. Various studies have both affirmed and refuted this claim, but recent meta-analyses have determined that there is no clear real-world link,[29] and a recent longitudinal fMRI study shows strong evidence against negative effects.[30] Certainly the sheer scale of computer games consumption today would require a proportional rise in the number of violent attacks, for which there is no evidence.

The argument that pornography has led to an increase in sexual violence is similarly controversial and difficult to measure.[31] Indeed a 2009 study showed an inverse relationship between online porn and reported rape in the United States.[32] While pornography may contribute to a culture where certain sexual practices may have negative consequences, there is no causal relationship. A report by the US National Online Resource Center on Violence Against Women states, "Pornography is neither a necessary nor sufficient condition for rape. . . . No one argues that if pornography disappeared that rape would disappear."[33]

As with porn, so with sex robots. It is difficult—if not impossible—to show that sex robots would lead to real-world violence against women, although the use of them may contribute to a larger-scale culture that is detrimental to women. But banning sex robots would not end rape. Recent reports of Santos's Samantha being "molested" at a trade show[34] may not indicate violence toward the robot. Indeed the use of the word "molested" is misleading: this is an inorganic object, not a person. Santos has countered claims by explaining that the doll was handled by thousands of people who had been told they were free to touch it.[35] Granting permission to touch a somewhat fragile object on display

will, naturally, result in damage to that object, even over a short period of time; this is well-known from museum interactions.[36]

While the common perception is the lonely, isolated, awkward, and unlovable man in his bedroom, their customers, say manufacturers, also include couples, widows, and those with disabilities. Abyss Creations have said that psychiatrists have used them in therapeutic treatment and that parents buy them for use by their socially excluded grownup children. There is no confusion among the owners that these dolls are human. They are human-like replicas, yes, and they are welcomed as that; they are given names, personalities, and backstories. They are, by and large, revered. The people who buy sex dolls report that they buy them for a number of different reasons. They are the collectors, hobbyists, admirers, lovers, enthusiasts, and addicts. Some want the feeling of company; others fetishize the sex. Some pose them and take photographs of them. Some are in human relationships; others are single. Some worship their dolls; others love them out of sentimentality. Some see them as sexual, others as romantic. All of them are aware that it is a doll, not a human, even if they choose to treat it as such.

Incidents of agalmatophilia—sexual attraction to a statue, doll, or mannequin—have been recounted in early Greek civilization. Alex Scobie and A.J. W. Taylor discuss the classical evidence for statue sex: eleven accounts from Ancient Greece and one from Italy.[37] Their work has been criticized as being unverifiable, although the early sexologist Iwan Bloch recorded a paraphilia called Venus Statuaria, or "statue rape."[38] Trudy Barber has studied communities where aficionados are sexually attracted to—and in some cases actually aim to become—sex dolls (and sex robots) in a fetish known as androidism. "There is a growing sub-culture," says Barber, "of people actually wishing to become robots and dolls explicitly through narcissist forms of sexual arousal and a cult of techno body fascism."[39]

Technofestishism is explored by Allison de Fren, who has written about technofetish communities such as alt.sex.fetish.robots (ASFR), after the early online Usenet group where the community initially gathered. She observes two groups within ASFR: those who desire an entirely artificial, built robot and those who desire a transformation from human to robot. Her research revealed that an ASFRian ethos was a feminization of objects: a clear implication and normalization of gender roles.[40] The crux of it, she writes, is the common interest in programmatic control. The ASFR community wiki describes it as "a human (typically female) who has been either willingly or unwillingly turned into any kind of inanimate object."

In my research with owners of the current generation of realistic sex dolls—those same people who are likely to buy the emerging sex robots—the emphasis is very much on a cherishing relationship.[41] This could, of course, be a

consequence of the cost of such dolls (from $5,000 and up for a RealDoll), but a genuine sense of consideration is apparent. Many owners of these dolls (often self-defined as "iDollators") are actually strongly invested in their care. They report that they choose a relationship with a doll for therapeutic reasons, including lack of intimacy with their human partner for physiological reasons, such as a spouse's chronic illness. Others prefer the safety of a doll following psychological problems resulting from the collapse of previous relationships.[42] The dolls are a proxy, to an extent. They either stand in for something—the wish for a Pygmalionesque transformation to a real woman—or they are worshipped and fetishized for what they are. Either way, these are not mistaken for a human-human relationship, nor do they replace it; they are more of a parallel. In which case, how important is it for these dolls and these doll's sex robot descendants to look human-like?

9.5. Appearance

Whether or not sexual companion robots should look human is currently a theoretical debate given that the only existing prototypes of sex robots resemble women. It's easy to identify two distinct branches in the development of sex technology: the sex toys, which have been around for millennia, and the sex robots, the twenty-first-century sex doll. Phallic-shaped objects date at least as far back as 28,000 BCE.[43] There are depictions of dildo use on Greek vases and accounts of sex toys in Greek texts, such as Aristophanes's *Lysistrata*, dating to 411 BCE. Medieval penitentials record punishments for their use. From the late nineteenth century onward these were augmented by the vibrator. Today there is a proliferation of smart sex toys made from new and interesting materials and capable of being programmed.

It wasn't until the latter part of the twentieth century that sex dolls became widely available commercially, but they have a much longer history. There are written references to sailors in the seventeenth century creating a *dame de voyage* or *dama de viaje*: women-shaped bundles of fabric and leather for sex-starved sailors to share. In Japan, soft, cushioned, fabric sex dolls have been used. These are referred to as *datch waifu*—Dutch wives—a reference to the dolls on the Dutch ships the Japanese merchants encountered in trading.[44] Bloch, in writing *The Sexual Life of Our Time* in 1909, refers to "clever mechanics who, from rubber and other plastic materials, prepare entire male or female bodies, which, as hommes or dames de voyage, subserve fornicatory purposes."[45]

Following the Western sexual revolution of the 1960s, the sex shops of the 1970s meant that sex toys with an undisguised purpose could be sold widely, although discretion was prevalent. The inclusion of sex toys in popular television

storylines such as HBO's *Sex and the City* saw a move toward the social acceptability of owning a sex toy, although the episode in which it featured presented it in a somewhat morally judgmental framework.[46] Interestingly, the vibrator shown in that series was a rabbit vibrator, the design of which originated in Japan, where obscenity laws led to the abstraction of phallic designs into something bright, cute, and colorful. The rabbit vibrator was a change in design, a move away from replicating human body parts. It showed that the form of sex toys could change, could be abstracted, could be optimized. Today there are hundreds of vibrators and dildos available, many of which look nothing like a penis.[47]

By contrast, penetrable sex toys—as differentiated from the full-body form of dolls—either originated much later or, more likely, we have lost evidence of them. Like the early sex dolls, substitutions were most probably made in the form of fabric or leather, mimicking orifices. Such materials mean those artifacts are unlikely to have survived in the archaeological record, although there are surviving *shunga* examples from Japan.[48] Artificial inflatable vaginas are mentioned in a 1922 catalog,[49] but it wasn't until the mid-1990s that the portable artificial vagina became a widespread commercial reality in the Western world when the Flashlight, a standalone artificial orifice, was widely marketed.[50]

The sex toy market is forecast to reach $30 billion worldwide by 2020. The past few years have seen new start-ups come onto the scene, capitalizing on advancements in materials and technology and an appetite for pleasure. In a 2018 study by the internet-based market research and data analytics company YouGov, 1,714 adults in the United Kingdom were surveyed about sex robots.[51] This was a follow-on from previous YouGov surveys in 2013 (in the United Kingdom) and 2017 (in the United States). Of the participants in the 2018 survey, 43.6% (748) were male and 56.4% (966) were female. Ages ranged from twenty to eighty-six. The median age was forty-nine. Following questions about sex toy ownership, the participants were asked "Would you consider having sex with a robot?" Of the 1,557 respondents to this question, 58.5% (1,003) said "No, definitely not," and 14.1% (242) said "No, probably not." In favor, 9.2% (157) said "Yes, probably"; 4% (69) said "Yes, definitely"; and 5% (86) responded "Don't know." Breakdown by gender showed that twice as a many men as women would consider having sex with a robot (152 men compared to 74 women). Of this group as a whole, 86% (194) said it was "very" or "somewhat" important that the robot looked human, whereas 10% (22) said that it didn't matter. Women were more likely than men to say that a humanoid appearance wasn't important.

Social norms are likely to play a role in reluctance to engage in sex with a robot, but these norms are dependent on what people perceive as a robot. In a study by Jessica M. Szczuka and Nicole C. Krämer, who investigated whether male participants would be likely to buy a sex robot, the robot examples shown to the men were videos of Sophia (by Hanson Robotics) and HRP-4C (Miim) by

the National Institute of Advanced Industrial Science and Technology.[52] Lynne Hall has remarked that domestic robots do not resemble maids or cleaners, and questions whether a human-like appearance is necessary for sexual activity.[53] This tallies with Clifford Nass et al.'s findings on computers as social actors, which showed that our social responses to machines are automatic and unconscious, and that the need for machines to have a realistic, human-like form is highly overrated.[54]

Recent work by Emily Cross et al. suggests that humans hold preconceived beliefs and expectations about robots and robotic behavior based on the robots' physical features but also on the people's own prior knowledge. Cross et al.'s findings indicate that "human knowledge about and attitudes towards robots will need to be optimized as much as a robot's physical form and motion."[55] We are primed by hundreds of years of robot stories to expect a human-like artificial lover. Current lifelike designs may therefore be attributable to expectations—a form of skeuomorphism where a sexual companion robot mimics the human form. However, we no longer expect our sex toys to represent realistic genitalia. I see parallels in interaction design where metaphors constrain innovation. Designers and developers choose to maintain the metaphor in line with familiarity for the user. The commonly cited example is the virtual calculator, which mirrors the limited form of the physical calculator despite being freed from physical conventions.[56]

The attempt to make hyperrealistic human-like robots may be doomed to failure if Masahiro Mori's postulation of the Uncanny Valley holds true.[57] However, empirical evidence on this theory is mixed, with some researchers reporting validity and others indicating no effect. Reasons for the possible revulsion when faced with not-quite-human robots are also unclear, ranging from mismatch between appearance and behavior, anticipation of sentience, and the taboo of death when confronted with something corpse-like.[58] Maya B. Mathur and David B. Reichling's work on studying any Uncanny Valley effects with eighty real-world robot faces indicated that "small faults in their [the robots'] humanness might send the social interaction tumbling."[59] Given that we are still a long way from convincingly human-like robots, combined with doubt over whether we will ever see sentient human-like robots, why are we still developing them in this form? While human traits may be desirable from an interaction standpoint, *realistically* human, that is, android or gynoid, seems an unachievable goal for now, and may indeed be the reason why the idea of sex with robots is still anathema to many.

The sex robots being developed today have a very specific female-gendered embodiment. The current appearance of prototype sex robots tends to be a reductive stereotype of the female form: hypersexualized and pornified bodies of

women, usually white, thin, and blonde. Krizia Puig's work on the lack of representation in the bodies of sex dolls and robots explores racial relationships of power and colonization. "The femmes of colour, the Crip ones, the queer ones, the poor ones, the ones from the third world—we are excluded from the future," writes Puig.[60] The prototype sex robots are built to serve the male gaze. They are artificial women for heterosexual men. In contrast, sex toys have been abstracted away from that and, because they are not a full humanoid form, are barely seen as gendered at all.

My observations are that sex toys have moved into a design-led phase where functionality has been refined and form is now paramount. By contrast, sex robots are still very much in an engineering-led phase, where functionality is key. The metaphor of the human still prevails: these are artificial products that still adhere to expected physical conventions. Sex robots have not yet moved into the design-led phase. New forms have not been explored. This transition from engineering to design is common. Software is one such example, especially in terms of user interfaces. Icons, for example, began as quite detailed attempts at realistic depictions of real-world items before moving to the flat style seen today. Design can be improved with the use of affordances, a term coined by James J. Gibson[61] to describe an object's possibilities for action, made physically possible by its properties; that is, perception drives action. Donald Norman[62] expanded on this with the term "perceived affordances," which refer to the actions users perceive to be possible when they encounter an object. Therefore, while human traits can be useful as perceived affordances, indicating how we should use or engage with an object, these too can be merely indicated rather than made explicit, if you'll pardon the pun.

The interesting future is the future in which the two separate paths of sex toys and sex dolls converge. Move away from the idea of the pornified fembot and we also move away from the perpetuation of objectification. Extending smart sex toy development into more embodied forms bridges the gap: if you want to design a sex robot, why not pick the features that could bring the greatest pleasure? A velvet or silk body, sensors and mixed genitalia; tentacles instead of arms? Two sex tech hackathons held at Goldsmiths, University of London in 2016 and 2017 explored innovative intimate technologies, including prototypes of sensory-stimulating blankets and hammocks that could hug.[63] While current prototype sex robots hinge on visual appearance and voice, a multisensory—or even a nonvisual approach—is also possible. (A study by Joon Huh et al.[64] showed brain activation areas of arousal in men using only an olfactory sexual stimulus, namely Chanel No. 5 perfume.) A wealth of multimodal, multisensory ideas emerge: perhaps a sex duvet, which can vibrate, squeeze, and purr, or a swarm of drones or robots moving around the body.

Perhaps more easily imagined—and therefore perhaps more palatable—is the artificially intelligent disembodied partner. Spike Jonze's 2013 film *Her* tells the story of a man who falls in love with his artificially intelligent operating system, Samantha. When viewed in the context of current sex robot design, *Her*'s Samantha is a clear aspiration: an AI that can tease and flirt and love; one that is always there for you and knows you from all your data. Back in the real world, the breach of the Ashley Madison dating site—a website aimed at helping married people meet secret partners—in 2015 showed that many of the men talking to women online on the site were actually talking to chatbots imitating women.[65] Abyss Creations' standalone Harmony app is a step on the way. There's no true intelligence in Harmony, but the concept is the same. The rapid spread of online dating and the use of the internet to form remote friendships and relationships is testament to the ease with which people can form attachments to individuals they have never met in real life. The disembodied lover is already here, as a human. And, like many human roles, it is one that could one day be under threat of automation.

9.6. Summary

The basic technology exists and the sex robot is now becoming a reality—and a commercial viability. However, the likelihood is that the current hyperrealistic gynoid will constitute a small and niche market, most likely bought by those currently buying companion sex dolls and those who seek novelty, such as the people using sex doll brothels. As such, the alleged threat they pose is very limited in scale. Beyond that, their appeal seems weak. Rather than a dedicated human-like, human-size sex robot, it seems slightly more believable that care and companion robots in the future could have their intrinsic purpose extended to include sexual functions. However, much more likely is the development of sex technology into increasingly embodied, abstracted forms, providing robotic, multisensory experiences. This does not negate ethical problems (indeed it raises new issues around data security, privacy, and user control and consent), but it reduces some of the more prevalent fears tied to the reductive portrayal of the female form. It also offers an innovative approach to enabling people to have an independent and fulfilling sex life where physiological, psychological, and discriminatory barriers currently exist. And it offers the opportunity to examine and explore the stigmatization of sex outside of monoheteronormative relationships. We have the chance to shape and explore technology, to make it more equal and diverse. We are not tied to one single form, nor should we be, and the inevitable growth of sex technology provides an opportunity to think ethically about its development.

Notes

1. Fiona Sturges, "The Sex Robots Are Coming: Seedy, Sordid—but Mainly Just Sad," *Guardian*, November 25, 2017, https://www.theguardian.com/tv-and-radio/2017/nov/25/sex-robots-are-coming-seedy-sordid-sad.
2. Libby Plummer and Cheyenne Macdonald, "Sex Robots Could Reveal Your Secret Perversions," *Mail Online*, December 21, 2016, http://www.dailymail.co.uk/sciencetech/article-4051008/Sex-robots-reveal-secret-perversions-Handing-intimate-data-privacy-risk-warns-expert.html.
3. Bryan Menegus, "Sex Robots May Literally Fuck Us to Death," *Gizmodo*, December 19, 2016, https://gizmodo.com/sex-robots-may-literally-fuck-us-to-death-1790276123.
4. John Danaher, "Should We Be Thinking about Robot Sex?," in *Robot Sex: Social and Ethical Implications*, ed. J. Danaher and N. McArthur (Cambridge, MA: MIT Press, 2017), 3–14.
5. J. Danaher and N. McArthur, eds., *Robot Sex: Social and Ethical Implications* (Cambridge, MA: MIT Press, 2017).
6. Genevieve Liveley, "Why Sex Robots Are Ancient History," *Conversation*, May 4, 2016, https://theconversation.com/why-sex-robots-are-ancient-history-58112.
7. Ibid.
8. See their website, https://realbotix.com/.
9. S. Santos and J. Vazquez, "The Samantha Project: A Modular Architecture for Modeling Transitions in Human Emotions," *International Robotics & Automation Journal* 3, no. 2 (2017): 275–280.
10. J. Nevett, "'It's the Next Big Thing': Male Sex Robots COMING in 2018 as Demand SKYROCKETS," *Daily Star*, January 6, 2018, https://www.dailystar.co.uk/news/latest-news/671766/male-sex-robots-dolls-2018-demand-skyrockets-realdoll-matt-mcmullen.
11. J. Wosk, *My Fair Ladies: Female Robots, Androids, and Other Artificial Eves* (New Brunswick, NJ: Rutgers University Press, 2015).
12. S. A. Sanders, B. J. Hill, W. L. Yarber, C. A. Graham, R. A. Crosby, and R. R. Milhausen, "Misclassification Bias: Diversity in Conceptualisations about Having 'Had Sex,'" *Sexual Health* 7, no. 1 (2010): 31–34.
13. M. Migotti and N. Wyatt, "On the Very Idea of Sex with Robots," In *Robot Sex: Social and Ethical Implications*, ed. J. Danaher and N. McArthur (Cambridge, MA: MIT Press, 2017), 15–27.
14. K. Richardson, "Sex Dolls and Sex Robots and Rape Culture," *Campaign Against Sex Robots* (blog), January 1, 2017. https://campaignagainstsexrobots.org/2017/01/01/sex-dolls-and-robots-and-rape-culture/.
15. N. McArthur, "The Case for Sexbots," in *Robot Sex: Social and Ethical Implications,* ed. J. Danaher and N. McArthur (Cambridge, MA: MIT Press, 2017), 31–46.
16. W. H. Masters, V. E. Johnson, and Reproductive Biology Research Foundation (U.S.), *Human Sexual Response* (New York: Bantam Books, 1986).
17. J. Bancroft, "The Endocrinology of Sexual Arousal," *Journal of Endocrinology* 186, no. 3 (2005): 411–27.

18. J. G. Pfaus, and L. A. Scepkowski, "The Biologic Basis for Libido," *Current Sexual Health Reports* 2, no. 2 (2005): 95–100.
19. F. Toates, "An Integrative Theoretical Framework for Understanding Sexual Motivation, Arousal, and Behavior," *Journal of Sex Research* 46 (2009): 168–93.
20. In this case, male and female bonobos, so not far removed from the human form.
21. H. A. Rupp, and K. Wallen, "Sex Differences in Response to Visual Sexual Stimuli: A Review," *Archives of Sexual Behavior* 37 (2008): 206.
22. K. Richardson, "The Asymmetrical 'Relationship': Parallels between Prostitution and the Development of Sex Robots," *Campaign Against Sex Robots* (blog), 2016, https://campaignagainstsexrobots.org/the-asymmetrical-relationship-parallels-between-prostitution-and-the-development-of-sex-robots/.
23. C. A. MacKinnon, "Prostitution and Civil Rights," *Michigan Journal of Gender & Law* 1 (1993): 13–31.
24. A. Dworkin, "Prostitution and Male Supremacy," *Michigan Journal of Gender & Law* 1 (1993): 1–12.
25. J. Danaher, B. Earp, and A. Sandberg, "Should We Campaign against Sex Robots?," in *Robot Sex: Social and Ethical Implications,* ed. J. Danaher and N. McArthur (Cambridge, MA: MIT Press, 2017), 47–72.
26. UN General Assembly, *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime* (Trafficking Protocol), November 15, 2000, Art. 3. https://www.unodc.org/documents/treaties/UNTOC/Publications/TOC%20Convention/TOCebook-e.pdf. Known as the Palermo Protocol, it states that *enforced* prostitution comes under the proposed definition of trafficking.
27. S. A. Law, "Commercial Sex: Beyond Decriminalization," *Southern California Law Review* 73 (1999): 523.
28. D. Levy, *Love and Sex with Robots: The Evolution of Human-Robot Relationships* (New York: Harper Collins, 2009).
29. C. J. Ferguson, "Do Angry Birds Make For Angry Children? A Meta-analysis of Video Game Influences on Children's and Adolescents' Aggression, Mental Health, Prosocial Behavior, and Academic Performance," *Perspectives on Psychological Science* 10, no. 5 (2015): 646–66.
30. S. Kühn, D. Kugler, K. Schmalen, M. Weichenberger, C. Witt, and J. Gallinat, "The Myth of Blunted Gamers: No Evidence for Desensitization in Empathy for Pain after a Violent Video Game Intervention in a Longitudinal fMRI Study on Non-Gamers," *Neurosignals* 26 (2018): 22–30.
31. M. S. C. Lim, E. R. Carrotte, and M. E. Hellard, "The Impact of Pornography on Gender-Based Violence, Sexual Health and Well-Being: What Do We Know?," *Journal of Epidemiology and Community Health* 70 (2016): 3–5.
32. C. J. Ferguson, and R. D. Hartley, "The Pleasure Is Momentary . . . the Expense Damnable? The Influence of Pornography on Rape and Sexual Assault," *Aggression and Violent Behavior* 14, no. 5 (2009): 323–29.
33. Robert Jensen, "Pornography and Sexual Violence," *VAWnet*, July 2004, https://vawnet.org/sites/default/files/materials/files/2016-09/AR_PornAndSV.pdf.

34. Greg Nichols, "Sex Robot Molested, Destroyed at Electronics Show," *ZDNet*, October 2, 2017, http://www.zdnet.com/article/sex-robot-molested-destroyed-at-electronics-show/.
35. "Falschmeldung geht um die Welt." *Der Standard*, September 29, 2017. https://derstandard.at/2000065016001/Falschmeldung-geht-um-die-Welt-Keine-Uebergriffe-auf-Sexroboter-in.
36. G. K. Talboys, *Museum Educator's Handbook*, 2nd ed. (Farnham, UK: Ashgate, 2005).
37. A. Scobie and A. J. W. Taylor, "Perversions Ancient and Modern: I. Agalmatophilia, the Statue Syndrome," *Journal of the History of the Behavioral Sciences* 11, no. 1 (1975): 49–54.
38. Iwan Bloch, *The Sexual Life of Our Time in Its Relations to Modern Civilization* (London: Rebman, 1909). This work has been digitized online as part of the Internet Archive, accessed February 2018, https://archive.org/details/b20442609.
39. T. A. Barber, "For the Loveof Artifice," paper presented at AISB 50th Symposium, London, April 2014.
40. de Fren, Allison, "Technofetishism and the Uncanny Desires of ASFR (alt.sex.fetish.robots)." *Science Fiction Studies* 36, no. 3 (2009): 404–40.
41. Kate Devlin, *Turned On: Science, Sex and Robots.* London: Bloomsbury, 2018.
42. Jenna Owsianik, "iDollator Culture: Inside the Minds of Men Who Love Dolls," *Future of Sex*, July 31, 2016, https://futureofsex.net/robots/idollator-culture-inside-minds-men-love-dolls/.
43. Jonathan Amos, "Ancient Phallus Unearthed in Cave," *BBC News*, July 25, 2005, http://news.bbc.co.uk/1/hi/sci/tech/4713323.stm;T. Taylor, *The Prehistory of Sex: Four Million Years of Human Sexual Culture* (London: Fourth Estate, 1996).
44. A. Ferguson, *The SexDoll: A History* (McFarland, 2010)..
45. Bloch, I., *The Sexual Life of Our Time in Its Relations to Modern Civilization*.
46. H. Lieberman, *Buzz: A Stimulating History of the Sex Toy* (New York: Pegasus Books, 2017).
47. Jon Millward, "Down the Rabbit-Hole: What One Million Sex Toy Sales Reveal about Our Erotic Tastes, Kinks and Desires," blog, September 8, 2014,http://jonmillward.com/blog/studies/down-the-rabbit-hole-analysis-1-million-sex-toy-sales/.
48. Cynthia A. Moya, "Artificial Vaginas and Sex Dolls: An Erotological Investigation" (PhD diss., Institute for Advanced Study of Human Sexuality, 2006).
49. H. N. Cary, *Erotic Contrivances: Appliances Attached to, or Used in Place of, the Sexual Organs* (privately printed, 1922).
50. US Patent US5782818A 1997-05-20 Shubin, S. A. Device for discreet sperm collection.
51. YouGov, A survey on public attitudes to sex robots, personal communication, January 2018.
52. J. M. Szczuka and N. C. Krämer, "Influences on the Intention to Buy a Sex Robot," in *Love and Sex with Robots: LSR 2016,* Lecture Notes in Computer Science, vol. 10237, ed. A. Cheok, K. Devlin, and D. Levy (Springer, 2017), 72–83.
53. L. Hall, "Sex with Robots for Love Free Encounters," in *Love and Sex with Robots: LSR 2016*, Lecture Notes in Computer Science, vol. 10237, ed. A. Cheok, K. Devlin, and D. Levy (Cham: Springer, 2017), 128–136.

54. C. Nass, J. Steuer, and E. R. Tauber, "Computers Are Social Actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM, 1994), 72–78.
55. E. S. Cross, R. Ramsey, R. Liepelt, W. Prinz, and A. F. de C. Hamilton, "The Shaping of Social Perception by Stimulus and Knowledge Cues to Human Animacy," *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, no. 1686 (2016): 20150075, https://royalsocietypublishing.org/doi/full/10.1098/rstb.2015.0075?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed&.
56. Y. Rogers, H. Sharp, and J. Preece, *Interaction Design: Beyond Human-Computer Interaction* (Hoboken, NJ: John Wiley & Sons, 2011), 45.
57. Masahiro Mori, "Bukimi no tani" [The uncanny valley], *Energy* 7, no. 4 (1970): 33–35; Masahiro Mori, "The Uncanny Valley," trans. Karl F. MacDorman and Norri Kageki, IEEE Spectrum, June 12, 2012, https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley.
58. See Stephanie Lay, "Uncanny Valley: Why We Find Human-like Robots and Dolls So Creepy," *Conversation*, November 10, 2015, https://theconversation.com/uncanny-valley-why-we-find-human-like-robots-and-dolls-so-creepy-50268 for an overview of research findings.
59. M. B. Mathur, and D. B. Reichling, "Navigating a Social World with Robot Partners: A Quantitative Cartography of the Uncanny Valley," *Cognition* 146 (2016): 22–32.
60. K. Puig, "The Synthetic Hyper Femme: On Sex Dolls, Fembots, and the Futures of Sex" (MA thesis, San Diego State University, 2017).
61. J. Gibson, *The Ecological Approach to Visual Perception* (New York: Psychology Press, 2015).
62. D. A. Norman, *The Design of Everyday Things* (New York: Doubleday, 1990).
63. H. Campbell, "Better Loving through Technology: A Day at the Sex-Toy Hackathon," *Observer*, December 10, 2017, https://www.theguardian.com/technology/2017/dec/10/better-loving-through-technology-sex-toy-hackathon.
64. J. Huh, K. Park, I. S. Hwang, S. I. Jung, H. J. Kim, T. W. Chung, and G. W. Jeong, "Brain Activation Areas of Sexual Arousal with Olfactory Stimulation in Men: A Preliminary Study Using Functional MRI," *Journal of Sexual Medicine* 5, no. 3 (2008): 619–25.
65. Annalee Newitz, "Ashley Madison Code Shows More Women, and More Bots," *Gizmodo*, August 31, 2015, https://gizmodo.com/ashley-madison-code-shows-more-women-and-more-bots-1727613924.

References

Amos, Jonathan. "Ancient Phallus Unearthed in Cave." *BBC News*, July 25, 2005. http://news.bbc.co.uk/1/hi/sci/tech/4713323.stm.

Bancroft, John. "The Endocrinology of Sexual Arousal." *Journal of Endocrinology* 186, no. 3 (2005): 411–27.

Barber, Trudy. A. "For the Love of Artifice." Paper presented at AISB 50th Symposium, London, April 2014.

Bloch, Iwan. *The Sexual Life of Our Time in Its Relations to Modern Civilization*. London: Rebman, 1909.

Campbell, Hayley. "Better Loving through Technology: A Day at the Sex-Toy Hackathon." *Observer*, December 10, 2017. https://www.theguardian.com/technology/2017/dec/10/better-loving-through-technology-sex-toy-hackathon.

Cary, Henry Nathaniel *Erotic Contrivances: Appliances Attached to, or Used in Place of, the Sexual Organs*. Privately printed, 1922

Cross, Emily S., Richard Ramsey, Roman Liepelt, Wolfgang Prinz, and Antonia F.C. Hamilton. "The Shaping of Social Perception by Stimulus and Knowledge Cues to Human Animacy." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, no. 1686 (2016): 20150075. https://royalsocietypublishing.org/doi/full/10.1098/rstb.2015.0075?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed&.

Danaher, John "Should We Be Thinking about Robot Sex?" In *Robot Sex: Social and Ethical Implications*, edited by J. Danaher and N. McArthur, 3–14. Cambridge, MA: MIT Press, 2017.

Danaher, John, Brian D. Earp, and Anders Sandberg. "Should We Campaign against Sex Robots?" In *Robot Sex: Social and Ethical Implications*, edited by John Danaher and Neil McArthur, 47–72. Cambridge, MA: MIT Press, 2017.

Danaher, John, and Neil McArthur, eds. *Robot Sex: Social and Ethical Implications.* Cambridge, MA: MIT Press, 2017.

de Fren, Allison. "Technofetishism and the Uncanny Desires of ASFR (alt.sex.fetish.robots)." *Science Fiction Studies* 36, no. 3 (2009): 404–40.

Devlin, Kate. *Turned On: Science, Sex and Robots.* London: Bloomsbury, 2018.

Dworkin, Andrea "Prostitution and Male Supremacy." *Michigan Journal of Gender & Law* 1 (1993): 1–12.

"Falschmeldung geht um die Welt." *Der Standard*, September 29, 2017. https://derstandard.at/2000065016001/Falschmeldung-geht-um-die-Welt-Keine-Uebergriffe-auf-Sexroboter-in.

Ferguson, Anthony *The Sex Doll: A History* (McFarland, 2010).

Ferguson, Christopher J. "Do Angry Birds Make For Angry Children? A Meta-analysis of Video Game Influences on Children's and Adolescents' Aggression, Mental Health, Prosocial Behavior, and Academic Performance." *Perspectives on Psychological Science* 10, no. 5 (2015): 646–66.

Ferguson, Christopher J., and Richard D. Hartley. "The Pleasure Is Momentary . . . the Expense Damnable? The Influence of Pornography on Rape and Sexual Assault." *Aggression and Violent Behavior* 14, no. 5 (2009): 323–29.

Gibson, James J. *The Ecological Approach to Visual Perception*. New York: Psychology Press, 2015.

Hall Lynne "Sex with Robots for Love Free Encounters." In *Love and Sex with Robots: LSR 2016*, Lecture Notes in Computer Science, vol. 10237, edited by A. Cheok, K. Devlin, and D. Levy. Cham: Springer, 2017: 128–136.

Huh, Joon, Kwangsung Park, In Sang Hwang, Seung Il Jung, Hyeong-Jung Kim, Tae-Woong Chung, and Gwang-Woo Jeong. "Brain Activation Areas of Sexual Arousal with Olfactory Stimulation in Men: A Preliminary Study Using Functional MRI." *Journal of Sexual Medicine* 5, no. 3 (2008): 619–25.

Jensen, Robert. "Pornography and Sexual Violence." *VAWnet*, July 2004. https://vawnet.org/sites/default/files/materials/files/2016-09/AR_PornAndSV.pdf.

Kühn, Simone, Dimitrij T. Kugler, Katharina Schmalen, Marcus Weichenberger, Charlotte Witt, and Jurgen Gallinat. "The Myth of Blunted Gamers: No Evidence for Desensitization in Empathy for Pain after a Violent Video Game Intervention in a Longitudinal fMRI Study on Non-Gamers." *Neurosignals* 26 (2018): 22–30.

Law, Sylvia A. "Commercial Sex: Beyond Decriminalization." *Southern California Law Review* 73 (1999): 523.

Lay, Stephanie. "Uncanny Valley: Why We Find Human-like Robots and Dolls So Creepy." *Conversation*, November 10, 2015. https://theconversation.com/uncanny-valley-why-we-find-human-like-robots-and-dolls-so-creepy-50268.

Levy, David. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Collins, 2009.

Lieberman, Hallie. *Buzz: A Stimulating History of the Sex Toy*. New York: Pegasus Books, 2017.

Lim, Megan S. C., Elise R. Carrotte, and Margaret E. Hellard. "The Impact of Pornography on Gender-Based Violence, Sexual Health and Well-Being: What Do We Know?" *Journal of Epidemiology and Community Health* 70 (2016): 3–5.

Liveley, Genevieve. "Why Sex Robots Are Ancient History." *Conversation*, May 4, 2016. https://theconversation.com/why-sex-robots-are-ancient-history-58112.

MacKinnon, Catharine A. 1993. "Prostitution and Civil Rights." *Michigan Journal of Gender & Law* 1 (1993): 13–31.

Masters, William H., Virginia E. Johnson, and Reproductive Biology Research Foundation (U.S.). *Human Sexual Response*. New York: Bantam Books, 1986.

Mathur, Maya B., and David B. Reichling. "Navigating a Social World with Robot Partners: A Quantitative Cartography of the Uncanny Valley." *Cognition* 146 (2016): 22–32.

McArthur, Neil. "The Case for Sexbots." In *Robot Sex: Social and Ethical Implications*, edited by J. Danaher and N. McArthur, 31–46. Cambridge, MA: MIT Press, 2017.

Menegus, Bryan. "Sex Robots May Literally Fuck Us to Death." *Gizmodo*, December 19, 2016. https://gizmodo.com/sex-robots-may-literally-fuck-us-to-death-1790276123.

Migotti, Mark and Nicole Wyatt. "On the Very Idea of Sex with Robots." In *Robot Sex: Social and Ethical Implications*, edited by J. Danaher and N. McArthur, 15–27. Cambridge, MA: MIT Press, 2017.

Millward, Jon. "Down the Rabbit-Hole: What One Million Sex Toy Sales Reveal about Our Erotic Tastes, Kinks and Desires." Blog. September 8, 2014. http://jonmillward.com/blog/studies/down-the-rabbit-hole-analysis-1-million-sex-toy-sales/.

Mori, Masahiro. "Bukimi no tani" [The uncanny valley]. *Energy* 7, no. 4 (1970): 33–35.

Mori, Masahiro. "The Uncanny Valley." Translated by Karl F. MacDorman and Norri Kageki. IEEE Spectrum, June 12, 2012. https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley.

Moya, Cynthia A. "Artificial Vaginas and Sex Dolls: An Erotological Investigation." PhD diss., Institute for Advanced Study of Human Sexuality, 2006.

Nass, Clifford, Jonathan Steuer, and Ellen R. Tauber. "Computers Are Social Actors." In Adelson, Beth, Susan Dumais, Judith Olson (eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. New York, NY: ACM, 1994.

Nevett, Joshua. "'It's the Next Big Thing': Male Sex Robots COMING in 2018 as Demand SKYROCKETS." *Daily Star*, January 6, 2018. https://www.dailystar.co.uk/news/

latest-news/671766/male-sex-robots-dolls-2018-demand-skyrockets-realdoll-matt-mcmullen.

Newitz, Annalee. "Ashley Madison Code Shows More Women, and More Bots," *Gizmodo*, August 31, 2015. https://gizmodo.com/ashley-madison-code-shows-more-women-and-more-bots-1727613924.

Nichols, Greg. "Sex Robot Molested, Destroyed at Electronics Show." *ZDNet*, October 2, 2017. http://www.zdnet.com/article/sex-robot-molested-destroyed-at-electronics-show/.

Norman, Donald A. *The Design of Everyday Things*. New York: Doubleday, 1990.

Owsianik, Jenna. "iDollator Culture: Inside the Minds of Men Who Love Dolls." *Future of Sex*, July 31, 2016. https://futureofsex.net/robots/idollator-culture-inside-minds-men-love-dolls/.

Pfaus, James G., and Lisa A. Scepkowski. "The Biologic Basis for Libido." *Current Sexual Health Reports* 2, no. 2 (2005): 95–100.

Plummer, Libby, and Cheyenne Macdonald. "Sex Robots Could Reveal Your Secret Perversions." *Mail Online*, December 21, 2016. http://www.dailymail.co.uk/sciencetech/article-4051008/Sex-robots-reveal-secret-perversions-Handing-intimate-data-privacy-risk-warns-expert.html.

Puig, Krizia. "The Synthetic Hyper Femme: On Sex Dolls, Fembots, and the Futures of Sex." MA thesis, San Diego State University, 2017.

Richardson, Kathleen. "The Asymmetrical 'Relationship': Parallels between Prostitution and the Development of Sex Robots." *Campaign Against Sex Robots* (blog), 2016. https://campaignagainstsexrobots.org/the-asymmetrical-relationship-parallels-between-prostitution-and-the-development-of-sex-robots/.

Richardson, Kathleen. "Sex Dolls and Sex Robots and Rape Culture." *Campaign Against Sex Robots* (blog), January 1, 2017. https://campaignagainstsexrobots.org/2017/01/01/sex-dolls-and-robots-and-rape-culture/.

Rogers, Yvonne, Helen Sharp, and Jennifer Preece. *Interaction Design: Beyond Human-Computer Interaction*. Hoboken, NJ: John Wiley & Sons, .

Rupp, Heather A., and Kim Wallen. "Sex Differences in Response to Visual Sexual Stimuli: A Review." *Archives of Sexual Behavior* 37 (2008): 206.

Sanders, Stephanie A., Brandon J. Hill, William L. Yarber, Cynthia A. Graham, Richard A. Crosby, and Robin R. Milhausen. "Misclassification Bias: Diversity in Conceptualisations about Having 'Had Sex.'" *Sexual Health* 7, no. 1 (2010): 31–34.

Santos, Sergi, and Javier Vazquez. "The Samantha Project: A Modular Architecture for Modeling Transitions in Human Emotions." *International Robotics & Automation Journal* 3, no. 2 (2017): 275–280.

Scobie, Alex, and A. J. W. Taylor. "Perversions Ancient and Modern: I. Agalmatophilia, the Statue Syndrome." *Journal of the History of the Behavioral Sciences* 11, no. 1 (1975): 49–54.

Sturges, Fiona. "The Sex Robots Are Coming: Seedy, Sordid—but Mainly Just Sad." *Guardian*, November 25, 2017. https://www.theguardian.com/tv-and-radio/2017/nov/25/sex-robots-are-coming-seedy-sordid-sad.

Szczuka Jessica. M., and Nicole C. Krämer. "Influences on the Intention to Buy a Sex Robot." In *Love and Sex with Robots: LSR 2016*, Lecture Notes in Computer Science, vol. 10237, edited by A. Cheok, K. Devlin, and D. Levy. Cham: Springer, 2017: 72–83.

Talboys, Graeme K. *Museum Educator's Handbook*. 2nd ed. Farnham, UK: Ashgate, 2005.

Taylor, Timothy. *The Prehistory of Sex: Four Million Years of Human Sexual Culture*. London: Fourth Estate, 1996.

Toates, Frederick. "An Integrative Theoretical Framework for Understanding Sexual Motivation, Arousal, and Behavior." *Journal of Sex Research* 46 (2009): 168–93.

UN General Assembly. *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime* (Trafficking Protocol). November 15, 2000. https://www.unodc.org/documents/treaties/UNTOC/Publications/TOC%20Convention/TOCebook-e.pdf.

Wosk, Julie. *My Fair Ladies: Female Robots, Androids, and Other Artificial Eves*. New Brunswick, NJ: Rutgers University Press, 2015.

Part III: Long-Term Impact of Superintelligence

10

Public Policy and Superintelligent AI

A Vector Field Approach

Nick Bostrom, Allan Dafoe, and Carrick Flynn

10.1. The Prospect of Radically Transformative AI

It has now become a widely shared belief that artificial intelligence (AI) is a general-purpose technology with transformative potential.[1] In this paper, we will focus on what is still viewed as a more controversial and speculative prospect: that of machine superintelligence—general artificial intelligence greatly outstripping the cognitive capacities of humans and capable of bringing about revolutionary technological and economic advances across a very wide range of sectors on timescales much shorter than those characteristic of contemporary civilization. We will not argue that this is a plausible or probable development;[2] rather, we will analyze some aspects of *what would follow* if radical machine superintelligence were in the cards for this century.

In particular, we focus on the implications of a machine intelligence revolution for governance and global policy. What would be a desirable approach to public policy under the assumption that we were approaching a transition to a machine superintelligence era? What general properties should one look for in proposals for how the world should manage the governance challenges that such a transition would bring with it?

We construe these questions broadly. Thus by "governance" we refer not only to the actions of states but also to transnational governance[3] involving norms and arrangements arising from AI technology firms, investors, NGOs, and other relevant actors, and to the many kinds of global power that shape outcomes.[4] And while ethical considerations are relevant, they do not exhaust the scope of the inquiry; we wish to include desiderata focused on the prudential interests of important constituencies as well as considerations of technical and political feasibility. We believe the governance challenges in the radical context that we focus on would in many respects be different from the issues that dominate discussions about more near-term AI developments.

It may be useful to say something briefly about the kinds of capabilities that we are imagining would be developed over the course of a transition to a

Nick Bostrom, Allan Dafoe, and Carrick Flynn, *Public Policy and Superintelligent AI* In: *Ethics of Artificial Intelligence.* Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0011

superintelligence era. As we picture the scenario, cheap, generally intelligent machines are developed that could substitute for almost all human labor, including scientific research and other inventive activity.[5] Early versions of machine superintelligence may quickly build more advanced versions, plausibly leading to an "intelligence explosion."[6] This acceleration of machine intelligence might drive other forms of technological progress, producing a plethora of innovations, such as in medicine and health, transportation, energy, education, and environmental sustainability. Economic growth rates would increase dramatically,[7] plausibly by several orders of magnitude.[8]

These developments will pose the challenge of making sure that AI is developed, deployed, and governed in a responsible and generally beneficial way. Some AI-related governance issues have begun to be explored, such as the ethics of lethal autonomous weapons;[9] AI-augmented surveillance;[10] fairness, accountability, and transparency in consequential algorithmic decisions;[11] and the design of domestic regulatory frameworks.[12] The transition to machine superintelligence, in particular, will pose substantial, even existential risks.[13] In the past several governmental bodies produced reports and announced national strategies on AI, including related governance challenges.[14]

For the purposes of this paper, the potential arrival of superintelligence this century, and other auxiliary claims about what this implies, can be regarded as *assumptions*. We do not pretend to offer sufficient evidence that they are plausible, but they help to define the hypothetical governance scenarios that we wish to analyze. A reader who is convinced that some claim is mistaken can view our analysis as a (possibly thought-provoking) intellectual exercise. Readers who attach some positive probability to these prospects might view our contribution as an effort to begin a conversation around the foundations for what could become the foremost policy issue later in this century: what a desirable approach to governance in a machine superintelligence era could look like.

10.2. A "Vector Field" Approach to Normative Analysis

Suppose that we optimistically conceive, in the most general terms, our overarching objective to be ensuring the realization of a widely appealing and inclusive near- and long-term future that ultimately achieves humanity's potential for desirable development, while being considerate to beings of all kinds whose interests may be affected by our choices. An ideal proposal for governance arrangements for a machine superintelligence world would then be one conducive to that end.

But what would this vague aspirational formulation mean in practice? Of course, there are many different views about the relative importance of various values and ethical norms, and there are many different actors (states, firms,

parties, individuals, NGOs, etc.) that have different ideological commitments and different preferences over how a future society should be organized and how benefits and responsibilities should be divided up. One way to proceed, in light of this multiplexity, would be to argue for one particular normative standard and seek to show how it is more attractive or rationally defensible than the alternatives. There is a rich literature, both in normative ethics and in wider political discourse, that attempts to do that. However, it is not our ambition in this paper to argue in favor of some particular fundamental ethical theory, normative perspective, social choice procedure, or political preference.

Another way to proceed would be to simply assume one particular normative standard, without argument, and then explore what follows from it regarding the particular matter at hand, then perhaps repeating this procedure for a range of different possible normative standards. This is also not what we will do here.

Instead, the approach we take in this paper is to attempt to be somewhat neutral toward many different commonly held normative views, ideologies, and private interests among influential actors. We do this by focusing on *the directional policy change*, from many possible evaluative standpoints, that is entailed by a set of special circumstances that can be expected to obtain in the scenario of radically transformative machine superintelligence that we outlined earlier.

In other words, we seek to sketch (metaphorically or qualitatively) a "vector field" of policy implications, which has relevance to a wide range of possible normative positions. For example, some political ideologies maintain that economic equality is a centrally important objective for public policy, while other ideologies maintain that economic equality is not especially important or that states have only very limited responsibilities in this regard (e.g., to mitigate the most extreme forms of poverty). The vector field approach might then attempt to derive directional policy change conclusions of a form that we might schematically represent as follows: "However much emphasis X you think that states ought, under present circumstances, to give to the objective of economic equality, there are certain special circumstances Y, which can be expected to hold in the radical AI context we described earlier, that should make you think that in *those* circumstances states should instead give emphasis $f_Y(X)$ to the objective of economic equality." The idea is that f here is some relatively simple function, defined over a space of possible evaluative standards or ideological positions. For instance, f might simply add a term to X, which would correspond to the claim "The emphasis given economic equality should be increased by a certain amount in the circumstances Y (according to all the ideological positions under consideration)." Or f might require telling a more complicated story, perhaps along the lines of "However much emphasis you give to economic equality as a policy objective under present circumstances, under conditions Y you should want to conceive of economic equality differently. Certain dimensions of economic

inequality are likely to become irrelevant and other dimensions are likely to become more important or policy-relevant than they are today." (We discuss equality-related issues in the section on allocation.)

This vector field approach is fruitful to the extent that there are some patterns in how the special circumstances *Y* impact policy assessments from different evaluative positions. If the prospect of radical AI had entirely different and idiosyncratic implications for every particular ideology or interest platform, then the function f would amount to nothing more than a lookup table. Policy analysis would then have to fall back to the ways of proceeding we mentioned earlier, that is, either trying to determine (or simply assuming) one uniquely correct or appropriate normative standard, or exploring a range of possible standards and investigating their policy implications separately.

We argue, however, that at least some interesting patterns can be found in f, and we strive to characterize some of them in what follows. We do this by first identifying several respects in which the prospect of superintelligent AI presents *special circumstances*—challenges or opportunities that are either unique to the context of such AI or are expected to present there in unusual ways or to unusual degrees. We then explain how these special circumstances have some relatively unambiguous implications for policy in the sense that there are certain policy properties that are far more important in these special circumstances (than they are in more familiar circumstances) for the satisfaction of many widely shared prudential and moral preferences. We express these especially relevant and important policy properties as a set of *desiderata*, or desirable qualities. The desiderata, which we arrange under four headings (efficiency, allocation, population, and process), are thus meant to express reasons for pushing policy in certain directions (relative to where the preferred policy point would be when we are operating outside of the special circumstances).

A strong proposal for the governance of advanced AI would ideally accommodate each of these desiderata to a high degree. There may exist additional desiderata that we have not identified here; we make no claim that our list is complete. Furthermore, a strong policy proposal should presumably also integrate many other normative, prudential, and practical considerations that are either idiosyncratic to particular evaluative positions or are not distinctive to the context of radical AI. Our contribution is to highlight some themes worth bearing in mind in further explorations of how we should approach governance and global policy challenges in light of the prospect of superintelligent AI.[15]

10.3. Efficiency

Under this heading we group desiderata that have to do with protecting or increasing the size of the pie that becomes available. An outcome would be

inefficient if it is Pareto inferior to some other possible outcome—for example, if it involves wasting resources, squandering opportunities for improvements, forfeiting achievable gains from mutually beneficial cooperation, and so forth. The desirability of greater efficiency may usually be taken for granted; however, there are some dimensions of efficiency that take on special significance in the context of a radical AI transformation. These include technical opportunity, AI risk, the possibility of catastrophic global coordination failures, and reducing turbulence.

10.3.1. Technological Opportunity

Machine superintelligence (of the type we are envisaging in this paper) would be able to expand the production-possibility frontier much further and far more rapidly than is possible under more normal circumstances. Superintelligent AI would be an extremely general-purpose technological advance, which could obviate most need for human labor and massively increase total factor productivity. In particular, such AI could make rapid progress in R&D and accelerate the approach to technological maturity.[16] This would enable the use of the fast outer realm of astronomical resources, including for settlement, which would become accessible to automated self-replicating spacecraft.[17] It would also open up a vast inner realm of development, making possible great improvements in health, lifespan, and subjective well-being, enriched life experiences, deeper understanding of oneself and others, and refinements in almost any aspect of being that we choose to cultivate.[18] Thus, in both the outward direction of extensive growth and the inward direction of intensive growth, dramatic progress could follow the development of superintelligence.

The surprisingly high ceiling for growth (and the prospect of a fast climb up to that ceiling) should make us think it especially important that this potential not be squandered. This desideratum has two aspects: (a) the inner and outer production-possibility frontiers should be pushed outward, so that Earth-originating life *eventually* reaches its full potential for realizing values, and (b) this progress should preferably occur *soon enough* that we (e.g., currently existing people, or any actors who are using these criteria to evaluate proposed AI paths) get to enjoy some of the benefits. The relative weight given to these two aspects will depend on an actor's values.[19]

Of particular note, there may be a level of technology that would allow human lifespan to become effectively unconstrained by biological aging and localized accidents—a level that would plausibly be reached not long after the emergence of superintelligence.[20] Consequently, for actors who care much about their own long-term survival (or the survival of their family or other existing people), the desirability of a path toward the development of superintelligent AI may depend

quite sensitively on whether it is likely to be fast enough to offer a chance for those people to have their lives saved by the AI transition.[21]

Even setting aside the possibility of life extension, how well existing people's lives go overall might fairly sensitively depend on whether their lives include a final segment in which they get to experience the improved standard of living that would be attained after a positive AI transition.

10.3.2. AI Risk

The avoidance of AI-induced destruction takes on special significance as a policy objective in the present context because it is plausible that the risk of such destruction—including especially extreme outcomes, such as human extinction—would not be, with the development of machine superintelligence, very small.[22] An important criterion for evaluating a proposed policy for long-term AI development is therefore how much quality-adjusted effort would be devoted to AI safety and supporting activities on that path. Relevant risk-reducing efforts may include, for example, pursuing basic research into scalable methods for AI control, encouraging AI builders to avail themselves of appropriate techniques, and more generally fostering conditions that ensure that the development of superintelligent AI is done with care and caution.

10.3.3. Possibility of Catastrophic Global Coordination Failures

Avoidance of catastrophic global coordination failures likewise has special significance in the present context, because such failures seem comparatively plausible there. Catastrophic coordination failure could arise in several ways.

Machine superintelligence could enable the discovery of technologies that make it easy to destroy humanity—for instance, by constructing some biotech- or nanotech-based "doomsday device," which, once invented, is cheap and easy to build. To stop ex ante or contain ex post the development of such an accessible doomsday device could require extreme and novel forms of global agreement, surveillance, restraint, and cooperation.

Coordination problems could lead to a risk-increasing AI technology race dynamic, in which developers throw caution to the wind as they vie to be the first to attain superintelligence.[23] A race dynamic could lead to reduced investment in safety research, reduced willingness to accept delays to install and test control methods, and reduced opportunities to rely on control methods that incur a significant computational cost or that otherwise hamper performance.

More generally, coordination failures could lead to various kinds of "races to the bottom" in the development and deployment of advanced AI. For instance, welfare provisions to protect the interests of artificial minds might be eroded in a hypercompetitive global economy in which jurisdictions that impose regulations against the mistreatment and exploitation of digital workers are competitively disadvantaged and marginalized. Evolutionary dynamics might also shape developments in undesirable directions and in ways that are impossible to avoid without effective global coordination.[24]

If technological developments increase the risk of catastrophic global coordination failure, then it becomes more important to develop options and mechanisms for solving those coordination problems. This could involve incremental work to improve existing global governance mechanisms and strengthen norms of cooperation.[25] It could also involve preferring development pathways that empower some actor with a decisive strategic advantage that could be used, if necessary, to stabilize the world when a substantial risk of existential coordination failure appears.[26]

10.3.4. Reducing Turbulence

The speed and magnitude of change in a machine intelligence revolution would pose challenges to existing institutions. Under highly turbulent conditions, preexisting agreements might fray and long-range planning become more difficult. This could make it harder to realize the gains from coordination that would otherwise be possible—both at the international level and within nations. At the domestic level, loss could arise from ill-conceived regulation being rushed through in haste, or well-conceived regulation failing to keep pace with rapidly changing technological and social circumstances. At the international level the risks of maladjustment are possibly even greater, as there are weaker governance institutions and less cultural cohesion, and it typically takes years or decades to conceive and implement well-considered norms, policies, and institutions. The resulting efficiency losses could take the form of temporary reductions in welfare or an increased risk of inferior long-term outcomes. Other things being equal, it is therefore desirable that such turbulence be minimized or well-managed.

10.3.5. Desiderata Related to Efficiency

From the preceding observations, we extract the following desiderata:

- *Expeditious progress.* This divides into two components: (a) Policies that lead with high probability to the eventual development of safe superintelligence

and its application to tapping novel sources of wealth; and (b) speedy AI progress, such that socially beneficial products and applications are made widely available in a timely fashion.

- *AI safety.* Techniques are developed that make it possible (without excessive cost, delay, or performance penalty) to ensure that superintelligent AI behaves as intended.[27] Also, the conditions during the emergence and early deployment of superintelligence are such as to encourage the use of the best available safety techniques and a generally cautious approach.
- *Conditional stabilization.* The development trajectory and the wider political context are such that *if* catastrophic global coordination failure would result in the absence of drastic stabilizing measures, *then* the requisite stabilization is undertaken in time to avert catastrophe. This might mean that there needs to be a feasible option (for some actor or actors) to establish a singleton or to institute a regime of intensive global surveillance or to strictly suppress the dissemination of dangerous technology or scientific knowledge.[28]
- *Nonturbulence.* The path avoids excessive efficiency losses from chaos and conflict. Political systems maintain stability and order, adapt successfully to change, and mitigate socially disruptive impacts.

10.4. Allocation

The distribution of wealth, status, and power is subject to perennial political struggle and dispute. There may not be much hope for a short section in a paper to add much novel insight to these century-old controversies. However, our vector field approach makes it possible for us to try to make some contribution to this subject without requiring us to engage substantially with the main issues under contention. Thus we focus here on identifying a few special circumstances which would surround the development of superintelligent AI, namely, risk externalities, reshuffling, the veil of ignorance, and cornucopia. These circumstances (we argue) should change the relative weight attached to certain policy considerations, norms, and values concerning allocation.[29]

10.4.1. Risk Externalities

As noted earlier, it has been argued that the transition to the machine intelligence era will be associated with some degree of existential risk. This is a risk to which all humans would be exposed, whether or not they participate in or consent to the project. A little girl in a village in Azerbaijan who has never heard about artificial intelligence would receive her share of the risk from the creation of machine

superintelligence. Fairness norms therefore require that she also receive some commensurate portion of the benefits if things turn out well. Consequently, to the extent that fairness norms form a part of the evaluation standard used by some actor, that actor should recognize as a desideratum that an AI development path provide for a reasonable degree of compensation or benefit-sharing to everybody it exposes to risk (a set that includes, at least, all humans who are alive at the time when the dangerous transition occurs).

Risk externalities appear often to be overlooked outside of the present (advanced AI) context too, so this desideratum could be generalized into a *Risk Compensation Principle*, which would urge policymaking aimed at the public good to consider arranging for those exposed to risk from another's activities to be compensated for the probabilistic harm they incur, especially in cases where full compensation if the actual harm occurs is either impossible (e.g., because the victim is dead or the perpetrator lacks sufficient funds or insurance coverage) or would not be forthcoming for other reasons.[30]

10.4.2. Reshuffling

Earlier we described the limitation of turbulence as an *efficiency*-related desideratum. Excessive turbulence could exact economic and social costs and, more generally, reduce the influence of human values on the future. But turbulence associated with a machine intelligence revolution could also have *allocational* consequences, and some of those point to additional desiderata.

Consider two possible allocational effects: *concentration* and *permutation*. By "concentration" we mean income or influence becoming more unequally distributed. In the limiting case, one nation, one organization, or one individual would own and control everything. By "permutation" we mean future wealth and influence becoming less correlated with present wealth and influence. In the limiting case, there would be zero correlation, or even an anticorrelation, between an actor's present rank (in, e.g., income, wealth, power, or social status) and that actor's future rank.

We do not claim that concentration or permutation will occur or that they are likely to occur. We claim only that they are salient possibilities and that they are *more* likely to occur to an extreme degree in the special circumstances that would obtain during a machine intelligence revolution than they are to occur (to a similarly extreme degree) under more familiar circumstances outside the context of advanced AI. Though we cannot fully justify this claim here, we can note, by way of illustration, some possible dynamics that could make this true. (1) In today's world, and throughout history, wage income is more evenly distributed than capital income.[31] Superintelligent AI, by strongly substituting for human

labor, could greatly increase the factor share of income received by capital.[32] All else being equal this would widen income inequality and thus increase concentration.[33] (2) In some scenarios, there are such strong first-mover advantages in the creation of superintelligence as to give the initial superintelligent AI, or the entity controlling that AI, a decisive strategic advantage. Depending on what that AI or its principal does with that advantage, the future could end up being wholly determined by this first-mover, thus potentially greatly increasing concentration. (3) When there is radical and unpredictable technological change, there might be more socioeconomic churn: some individuals or firms turn out to be well positioned to thrive in the new conditions or make lucky bets, and reap great rewards; others find their human capital, investments, and business models quickly eroding. A machine intelligence revolution might amplify such churn and thereby produce a substantial degree of permutation.[34] (4) Automated security and surveillance systems could make it easier for a regime to sustain itself without support from wider elites or the public. This would make it possible for regime members to appropriate a larger share of national output and to exert more fine-grained control over citizens' behavior, potentially greatly increasing the concentration of wealth and power.[35]

To the extent that one disvalues (in expectation) concentrating or permuting shifts in the allocation of wealth and power—perhaps because one places weight on some social contract theory or other moral framework that implies that such shifts are bad, or simply because one expects to be among the losers—one should thus regard continuity as a desideratum.[36]

10.4.3. Veil of Ignorance

At the present point in history, important aspects of the future remain at least partially hidden behind a veil of ignorance.[37] Nobody is sure when advanced AI will be created, where, or by whom (although, admittedly, some locations seem less probable than others). With most actors having fairly rapidly diminishing marginal utility in wealth, and thus risk-aversion in wealth, this would make it generally advantageous if an insurance-like scheme were adopted that would redistribute some of the gains from machine superintelligence.

It is also plausible that typical individuals have fairly rapidly diminishing marginal utility in power. For example, most people would much rather be certain to have power over one life (their own) than have a 10% chance of having power over the lives of ten people and a 90% chance of having no power. For this reason, it would also be desirable for a scheme to preserve a fairly wide distribution of power, at least to the extent of individuals retaining a decent degree of control over their own lives and their immediate circumstances (e.g., by having

some amount of guaranteed power or some set of inalienable rights). There is also international agreement that individuals should have substantial rights and power.[38]

10.4.4. Cornucopia

The transition to machine superintelligence could bring with it a bonanza of vast proportions. For example, Hanson estimates that cheap human-level machine intelligence would plausibly suffice to increase world GDP by several orders of magnitude within a few years after its arrival.[39] The ultimate magnitude of the economic potential that might be realized via machine superintelligence could be astronomical.[40]

Such growth would make it possible, using a small fraction of GDP, to nearly max out many values that have diminishing returns in resources (over reasonable expenditure brackets).

Suppose, for example, that the economy were to expand to the level where spending 5% of GDP would suffice to provide the entire human population with a guaranteed basic annual income of $40,000 plus access to futuristic-quality healthcare, entertainment, and other marvelous goods and services.[41] The case for adopting such a policy would then seem stronger than is the case today for instituting a guaranteed basic income, at a time when a corresponding policy would yield far less generous benefits, require the redistribution of a larger percentage of GDP, and threaten to dramatically reduce the supply of labor.

Similarly, if one state became so wealthy that by spending just 0.1% of its GDP on foreign aid, it could give everybody around the world an excellent quality of life (where there would otherwise be widespread poverty), then it would be especially desirable that the rich state does have at least that level of generosity. Whereas for a poor state, it does not much matter whether it gives 0.1% of GDP or it gives nothing—in neither case is the sum enough to make much difference—for an *extremely* rich state it could be crucially important that it gives 0.1% rather than 0%. In a really extreme case, it might not matter so much whether a superrich state gives 0.1% or 1% or 10%: the key thing is to ensure that it does not give 0%.

Or consider the case of a trade-off that a social planner faces between the value of animal welfare and the desire of many human consumers to have meat in their diet. Let us suppose that the planner cares mostly about human consumer preferences, but also cares a little about animal welfare. At a low level of GDP, the planner might choose to allow factory farming because it lowers the cost of meat. As GDP rises, however, there comes a point when the planner introduces legislation to discourage factory farming. If the planner did not care *at all* about animal

welfare, that point would never come. With GDP at modest levels, a planner that cares a lot about animal welfare might introduce legislation, whereas a planner that cares only a little about animal welfare might permit factory farming. But if GDP rises to sufficiently extravagant levels, then it might not matter how much the planner cares about animal welfare, so long as she cares *at least a tiny little bit*.[42]

Thus it appears that whereas today it may be more important to encourage higher rather than lower levels of altruism, in a cornucopia scenario the most important thing would not be to maximize the expected amount of altruism but to minimize the probability that the level of altruism ends up being zero. In cornucopian scenarios, we might say, it is especially desirable that epsilon-magnanimity prevails. More would be nice, and is supported by some of the other desiderata mentioned in this paper, but there is a special desirability to have a guaranteed floor that is significantly above the zero level.

More generally, it seems that if there are resource-satiable values that have a little support (and no direct opposition) and that compete with more strongly supported values only via resource constraints, then it would be desirable that those resources-satiable weaker values get at least some small fraction of the resources available in a cornucopian scenario such that they would indeed be satisfied.[43]

A future in which epsilon-magnanimity is ensured seems intuitively preferable. There are several possible ways to ground this intuition. (1) It would rank higher in the preference ordering of many current stakeholders, especially stakeholders that have resource-satiable interests that are currently dominated because of resource constraints. (2) It would be a wise arrangement in view of normative uncertainty: if dominant actors assign some positive probability to various resource-satiable values or moral claims being true, and it would be trivial to give those values their due in a cornucopian scenario, then a "moral parliament"[44] or other framework for dealing with normative uncertainty may favor policies that ensure an epsilon-magnanimity future. (3) Actors who have a desire or who recognize a moral obligation to be charitable or generous (or more weakly, to not be a complete jerk) may have reason to make a special effort to ensure an epsilon-magnanimous future.[45]

10.4.5. Desiderata Related to Allocation

These observations suggest that the assessment criteria with regard to the allocational properties of long-term AI-related outcomes include the following:

- **Universal benefit.** Everybody who is alive at the transition (or who could be negatively affected by it) gets some share of the benefit, in compensation for the risk externality to which they were exposed.

- **Epsilon-magnanimity.** A wide range of resource-satiable values (ones to which there is little objection aside from cost-based considerations) are realized if and when it becomes possible to do so using a minute fraction of total resources. This may encompass basic welfare provisions and income guarantees to all human individuals. It may also encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion.[46]
- **Continuity.** The path affords a reasonable degree of continuity such as to (i) maintain order and provide the institutional stability needed for actors to benefit from opportunities for trade behind the current veil of ignorance, including social safety nets; and (ii) prevent concentration and permutation from being unnecessarily large.

10.5. Population

Under this heading we assemble considerations pertaining to the creation of new beings, especially digital minds that have moral status or that otherwise matter to policymakers for noninstrumental reasons.

Digital minds can differ in fundamental ways from familiar biological minds. Distinctive properties of digital minds may include being easily and rapidly copyable, being able to run at different speeds, being able to exist without visible physical shape, having exotic cognitive architectures, having nonanimalistic motivation systems or perhaps precisely modifiable goal content, being exactly repeatable when run in a deterministic virtual environment, and having potentially indefinite lifespans.

The creation of beings with these and other novel properties would have complex and wide-ranging consequences for practical ethics and public policy. While most of these consequences must be set aside for future investigations, we can identify two broad areas of concern: the interests of digital minds and population dynamics.[47]

10.5.1. The Interests of Digital Minds

Advances in machine intelligence may create opportunities for novel categories of wrongdoing and oppression. The term "mind crime" has been used to refer to computations that are morally problematic because of their intrinsic properties, independently of their effects on the outside world, for example, because they instantiate sentient minds that are mistreated.[48] The issue of mind crime may arise well before the attainment of human-level or superintelligent AI.

Some nonhuman animals are widely assumed to be sentient and to have degrees of moral status. Future AIs, possessing similar sets of capabilities or cognitive architectures, may plausibly have similar degrees of moral status. Some AIs that are functionally very different from any animal might also have moral status.

Digital beings with mental life might be created on purpose, but they could also be generated inadvertently. In machine learning, for example, large numbers of agents are often generated during training procedures—many semifunctional versions of a reinforcement learner are created and pitted against one another in self-play; many fully functional agent instantiations are created during hyperparameter sweeps; and so forth. It is quite unclear just how sophisticated artificial agents can become before attaining some degree of morally relevant sentience—or before we can no longer be confident that they possess no such sentience.

Several factors combine to mark the possibility of mind crime as a salient special circumstance of advanced developments in AI. One is the novelty of sentient digital entities as moral patients. Policymakers are unaccustomed to taking into account the welfare of digital beings. The suggestion that they might acquire a moral obligation to do so might appear to some contemporaries as silly, just as laws prohibiting cruel forms of recreational animal abuse once appeared silly to many people.[49] Related to this issue of novelty is the fact that digital minds can be invisible, running deep inside some microprocessor, and that they might lack the ability to communicate distress by means of vocalizations, facial expressions, or other behaviors apt to elicit human empathy. These two factors, the novelty and potential invisibility of sentient digital beings, combine to create a risk that we will acquiesce in outcomes that our own moral standards, more carefully articulated and applied, would have condemned as unconscionable.

Another factor is that it can be unclear what constitutes mistreatment of a digital mind. Some treatments that would be wrongful if applied to sentient biological organisms may be unobjectionable when applied to certain digital minds that are constituted to interpret the stimuli differently. These complications increase when we consider more sophisticated digital minds (e.g., humanlike digital minds) that may have morally considerable interests in addition to freedom from suffering and interests such as survival, dignity, knowledge, autonomy, creativity, self-expression, social belonging, and political participation.[50] The combinatorial space of different kinds of mind with different kinds of morally considerable interests could be hard to map and hard to navigate.

A fourth factor, amplifying the other three, is that it may become inexpensive to generate vast numbers of digital minds. This will give more agents the power to inflict mind crime and to do so at scale. With high computational speed or parallelization, a large amount of suffering could be generated in a small amount of wall clock time. It is plausible that the vast majority of all minds that will ever

have existed will be digital. The welfare of digital minds, therefore, may be a principal desideratum in selecting an AI development path for actors who either place significant weight on ethical considerations or who for some other reason strongly prefer to avoid causing massive amounts of suffering.

10.5.2. Population Dynamics

Several concerns flow from the possibility of introducing large numbers of new beings, especially when these new beings possess attributes associated with personhood. Some of these concerns relate to the possibility of mind crime, which we discussed in the previous subsection, but other concerns pertain even if we assume that no mind crime takes place. One special circumstance that is relevant here is that, with digital replication rates, population numbers could change extremely rapidly. An active population policy, with appropriate arrangements put in place in advance, may be necessary to forestall Malthusian outcomes (where average income falls to close to subsistence level) and other bad results.

Consider the system of child support common in developed countries. Individuals are free to have as many children as they are able to create; the state steps in to support children whose parents fail to provide for them. With digital beings, this arrangement is obviously unsustainable. If parents were able to create arbitrary numbers of children and there is persistent variation in willingness to do so, this system would quickly collapse. It is true that over longer timescales, Malthusian concerns will arise for biologically reproducing persons as well, as evolution acts on human dispositions to select for types that take advantage of modern prosperity to generate larger families.[51] For digital minds, however, the onset of a Malthusian condition could be abrupt.[52]

Societies would thus confront a dilemma: *either* accept population controls, requiring would-be procreators to meet certain conditions before being allowed to create new beings, *or* accept the risk that vast numbers of new beings will be given only the minimum amount of resources required to support their labor, while being worked as hard as possible and terminated as soon as they are no longer cost-effective. Of these options, the former seems preferable, especially if it should turn out that the typical mental state of a maximally productive worker in the future economy is wanting in positive affect or other desirable attributes.[53]

Malthusian outcomes is one example of how population change could create problematic conditions on the ground. Another is the undermining of democracy that can occur if the sizes of different demographics are subject to manipulation. Suppose that some types of digital beings obtain voting rights, on a one-person-one-vote basis. Such an enfranchisement might occur because humans give some class of digital minds voting rights for moral reasons or

because a large population of high-performing digital minds is effective at exerting political influence. This new segment of the electorate could then be rapidly expanded by means of copying, to the point where the voting power of the original human block is decisively swamped.[54] All copies from a given template may share the same voting preferences as the original, creating an incentive for digital beings to create numerous copies of themselves—or of more resource-efficient surrogates designed to share the originator's voting preferences and to satisfy eligibility requirements—in order to increase their political influence. This would present democratic societies with a trilemma: they could *either* (i) deny equal votes to all persons (excluding from the franchise digital minds that are functionally and subjectively equivalent to a human being); *or* (ii) impose constraints on creating new persons (of the type that would qualify for suffrage if they were created); *or* (iii) accept that voting power becomes proportional to ability and willingness to pay to create voting surrogates, resulting in both economically inefficient spending on such surrogates and the political marginalization of those who lack resources or are unwilling to spend them on buying voting power.[55]

10.5.3. Desiderata Related to Population

A full accounting of how the special circumstances of advanced AI should affect population policy would require a far more fine-grained analysis, but the preceding discussion lets us identify two broad desiderata:

- **Mind crime prevention.** Advanced AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized.
- **Population policy.** Procreative choices, concerning what new beings to bring into existence, are made in a coordinated manner and with sufficient foresight to avoid unwanted Malthusian dynamics and political erosion.

10.6. Process

The previous desiderata are expressed in terms of features of *outcomes*. We can also formulate desiderata in terms of properties that we want to pertain to the *process* through which the future gets determined. Here we point to three special circumstances with implications for governance that may plausibly obtain around the emergence of superintelligent AI: novelty, depth, and technical challenge of the policy context; pace of events; and the undermining of prevailing principles and norms.

10.6.1. Epistemic Challenge (Novelty, Depth, and Technicality)

The context of a machine intelligence revolution would place unusual epistemic demands on the policymaking process.

First, an impending or occurring machine intelligence revolution would entail an exceptionally large shift in the policymaking context. This means that many customary assumptions—such as are embedded in institutional arrangements, mental habits, and cultural norms—may become inapplicable. This would place a premium on being able to see the situation afresh by thinking things through from first principles or by being able to draw on an extremely wide and diverse experience base.

Second, and relatedly, the challenges confronting decision-makers in this context may come to involve fundamental worldview questions of a type that impinge on deep empirical, philosophical, strategic, or religious issues, and which are often clouded in uncertainty or controversy. This points to a special need for *wisdom*. Although difficult to operationalize, we take wisdom to mean the ability to reliably get the most important things at least approximately right. Wisdom involves a kind of robustly good judgment, well-calibrated degrees of belief, and a knack for finding a sensible path through a tricky and confusing situation, keeping the bigger picture in mind. In particular, it involves having a sufficient degree of epistemic humility to recognize the limits of one's knowledge and to be able to change one's mind, even about quite fundamental things, rather than persisting indefinitely with some catastrophically mistaken plan.

Third, since we are postulating a decision-making context in which an absolutely critical factor is a technological invention, there is a greater-than-usual premium on being able to understand technology—especially AI technology—and form appropriate expectations about its attributes and potentialities. To some extent, this desideratum might be satisfied by bringing in appropriate technical experts to advise policymakers. But the governance mechanism as a whole needs to be such that the right experts are selected, listened to, and understood. And other things equal, a decision-maker who is ignorant of science and technology and incapable of following a mathematical or technical argument, and is thus reduced to conceptualizing the AI technology as a black box about which different accredited scientific experts make cryptic and sometimes conflicting edicts, is probably at a disadvantage compared to a decision-maker who is able to form a reasonable mechanism-level understanding of the technology under consideration.

10.6.2. Pace

In many scenarios, events of world-historic consequence would be unfolding at an unusually fast pace during the transition to machine superintelligence. This

suggests that it may be more important than it normally is for governance processes to be able to move rapidly and decisively, to stay ahead of events. In particular, it may be desirable that the development of superintelligent AI takes place in a governance context in which it is possible to make constitutional changes quickly and to decide and impose global governance arrangements on timescales much shorter than those typically associated with negotiating, ratifying, and implementing multinational treaties.

10.6.3. Undermining

There are various ways in which the context of a machine intelligence revolution may present special opportunities for principles and norms to be undermined or for existing power structures to be usurped. We touched on some of these in our discussion about "reshuffling," in terms of how social outcomes might be subject to extreme degrees of permutation or concentration of wealth and influence. But we can also approach these matters from a process-oriented perspective.

Consider principles such as legitimacy, consent, political participation, and accountability. These are widely thought to be desirable attributes for governance systems and policymaking processes. Yet the special circumstances of a machine intelligence revolution could undermine these principles in various ways.

Take, for example, the idea of voluntary consent, a hugely important principle that regulates many interactions between both individuals and states. Many things that it would be morally wrong or illegal to do to an individual without her consent are entirely unobjectionable if done with her consent. The same holds for many possible interactions between corporate entities or states: it very often makes a world of difference whether something is taken or imposed by force, or voluntarily agreed to. Yet consider how this central role given to consent could be undermined in the context of advanced AI, if it becomes possible to construct a "super-persuader," a system that has the ability, through extremely skillful use of argumentation and rhetoric, to persuade almost any human individual or group (unaided by similarly powerful AI) of almost any position or to get them to accept almost any deal. Should it be possible to create such a super-persuader, then it would be inappropriate to continue to rely on consent as a near-sufficient condition for many types of transaction to be morally and legally unobjectionable. In a world with super-persuaders, there would need to be stronger protections to safeguard human interests, analogous to the extra safeguards currently in place to protect the interests of certain classes of vulnerable individuals, such as children and adults with cognitive impairments. Perhaps consent should be regarded as valid only if the human counterparty had access to a qualified AI

advisor or if the transaction were approved by an "AI guardian" assigned to the human actor to protect her from exploitation.

For another example, consider the norm of political participation. This norm might be justified on several different grounds. On the one hand, it could provide an epistemic benefit by including more information and a broader range of perspectives into the decision-making process. On the other hand, it could also be a way of ensuring that many different interests and preferences are reflected in the decisions that are made. And on the prehensile tail, political participation could be regarded as an intrinsic good, to be valued independently of any contribution it makes to producing decisions that better serve all the interests concerned.[56] These three justifications may need to be reevaluated in the context of superintelligent AI. For instance, it is possible that the epistemic value of letting political decisions be influenced by many human opinions would be reduced or eliminated if superintelligent AI were sufficiently epistemically superior to humans and able to discern and integrate independently all the scraps of evidence and insight that a distributed human epistemic community would have been able to supply. It is also conceivable that advanced AI would enable the construction of a mechanism that does not require the continual input of human preference articulations in order to factor those preferences into the decisions that are being made; maybe a superintelligent AI could learn a preference function that already anticipates the existing distribution of human preferences and the shifts in those preferences that will occur over time, or the AI might be able to infer this from observing other kinds of human behavior. The supposed intrinsic value of political participation might remain intact even if the two instrumental justifications were to disappear, or perhaps it would come to be seen as quaint and perverse to want to participate in political affairs after it becomes clear that one's interventions serve only to make political outcomes worse (for both one's own interests and those of the wider society).

The purpose of these two examples is not to advance specific claims about consent or political participation in an era of superintelligent AI, but to illustrate a more general point: that there are various principles and norms, which are currently deeply entrenched and often endorsed without qualification, that would need to be examined afresh in a context of radical AI.[57] Some of these norms and principles may have to be abandoned in that context; others may need to be reinterpreted and reformulated; and yet others may need to be safeguarded with greater than usual vigilance. This points to a general desideratum on governance processes in this context, namely, that they be capable of leading to appropriate adaptation of relevant norms and principles.[58]

10.6.4. Desiderata Related to Process

From the preceding observations, we derive a set of desiderata pertaining to the governance processes by which policy is decided in the context of superintelligent AI:

- **First-principles thinking, wisdom, technical understanding.** The transition to superintelligent AI is governed by some agency (individual or collective, centralized or distributed) that is able to effectively integrate uncommon levels of first-principles thinking, wisdom, and technical understanding into its decision-making.
- **Speed and decisiveness.** Development and deployment of superintelligent AI is done in a political context in which there exists a capacity for rapid decision-making and decisive global implementation (or, alternatively, a capacity to moderate the pace of developments so as to allow slower decision-making and coordination processes to be effective).
- **Adaptability.** Superintelligent AI is deployed in a sociopolitical context in which rules, principles, norms, and laws can be adapted as appropriate to fit the novel circumstances.

10.7. Summary

We have drawn attention to a number of special circumstances that may surround the development and deployment of superintelligent AI, circumstances that present distinctive challenges for governance and global policy. Using a "vector field" approach to normative analysis, we sought to extract directional policy implications from these special circumstances. We characterized these implications as a set of desiderata—traits of future policies, governance structures, or decision-making contexts that would, by the standards of a wide range of key actors, stakeholders, and ethical views, enhance the prospects of beneficial outcomes in the transition to a machine intelligence era. These desiderata (which we do not claim to be exhaustive) are summarized in Table 10.1.

The desiderata in Table 10.1 help establish criteria by which concrete policy proposals for the governance of advanced AI could be evaluated. By "policy proposals" we refer not only to official government documents but also to plans and options developed by private actors who take an interest in long-term AI developments. The desiderata, therefore, are also relevant to some corporations, research funders, academic or nonprofit research centers, and various other organizations and individuals.

Table 10.1 Special Circumstances Expected to Be Associated with the Transition to a Machine Intelligence Era (Left Column) and Corresponding Desiderata for Governance Arrangements (Right Column)

	Efficiency
Technological opportunity AI risk Possibility of catastrophic global coordination failures Reducing turbulence	Expeditious progress. This divides into two components: (a) Policies that lead with high probability to the eventual development of safe superintelligence and its application to tapping novel sources of wealth; and (b) speedy AI progress, such that socially beneficial products and applications are made widely available in a timely fashion. AI safety. Techniques are developed that make it possible (without excessive cost, delay, or performance penalty) to ensure that superintelligent AI behaves as intended. Also, the conditions during the emergence and early deployment of superintelligence are such as to encourage the use of the best available safety techniques and a generally cautious approach. Conditional stabilization. The development trajectory and the wider political context are such that if catastrophic global coordination failure would result in the absence of drastic stabilizing measures, then the requisite stabilization is undertaken in time to avert catastrophe. This might mean that there needs to be a feasible option (for some actor or actors) to establish a singleton or to institute a regime of intensive global surveillance or to strictly suppress the dissemination of dangerous technology or scientific knowledge. Nonturbulence. The path avoids excessive efficiency losses from chaos and conflict. Political systems maintain stability and order, adapt successfully to change, and mitigate socially disruptive impacts.
	Allocation
Risk externalities Reshuffling Veil of ignorance Cornucopia	Universal benefit. Everybody who is alive at the transition (or who could be negatively affected by it) gets some share of the benefit, in compensation for the risk externality to which they were exposed. Epsilon-magnanimity. A wide range of resource-satiable values (ones to which there is little objection aside from cost-based considerations) are realized if and when it becomes possible to do so using a minute fraction of total resources. This may encompass basic welfare provisions and income guarantees to all human individuals. It may also encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion. Continuity. The path affords a reasonable degree of continuity such as to (i) maintain order and provide the institutional stability needed for actors to benefit from opportunities for trade behind the current veil of ignorance, including social safety nets; and (ii) prevent concentration and permutation from being unnecessarily large.

Continued

Table 10.1 *Continued*

	Population
Interests of digital minds	Mind crime prevention. Advanced AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized. Population policy. Procreative choices, concerning what new beings to bring into existence, are made in a coordinated manner and with sufficient foresight to avoid unwanted Malthusian dynamics and political erosion.
Population dynamics	
	Process
Epistemic challenge (novelty, depth, and technicality)	First-principles thinking, wisdom, technical understanding. The transition to superintelligent AI is governed by some agency (individual or collective, centralized or distributed) that is able to effectively integrate uncommon levels of first-principles thinking, wisdom, and technical understanding into its decision-making. Speed and decisiveness. Development and deployment of advanced AI is done in a political context in which there exists a capacity for rapid decision-making and decisive global implementation (or, alternatively, a capacity to moderate the pace of developments so as to allow slower decision-making and coordination processes to be effective). Adaptability. Superintelligent AI is deployed in a sociopolitical context in which rules, principles, norms, and laws can be adapted as appropriate to fit the novel circumstances.
Pace	
Undermining	

The development of concrete proposals that might satisfy these desiderata is a task for further research. Such concrete proposals would probably need to be relativized to specific actors, since the best way to comport with the general considerations we have identified will depend on the capacities, resources, and political constraints of the actor to whom the proposal is directed. Furthermore, specific actors may also have additional idiosyncratic preferences that are not fully captured by our vector field analysis but that must be accommodated in order for a policy proposal to stand a chance of gaining acceptance.[59]

Notes

1. Among many examples of reports of this kind, see Darrell M. West and John R. Allen, *How Artificial Intelligence Is Transforming the World,* Brookings Institution, April 24, 2018, https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.
2. Nevertheless it is worth noting that many AI researchers take the prospect of superintelligence in this century seriously. Indeed, within the machine learning community,

the majority view is that it is more likely than not that human-level machine intelligence will be developed by 2050 (Vincent C. Müller and Nick. Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* [Geneva: Springer International, 2016], 553–71) or 2060 (Katja Grace et al. "When will AI exceed human performance? Evidence from AI experts." *Journal of Artificial Intelligence Research*, 62 (2018): 729-754.), and that it is likely (75%) that superintelligence will be developed within thirty years after.

3. Thomas Hale and David Held, *Handbook of Transnational Governance* (Cambridge, UK: Polity Press, 2011).
4. Michael Barnett and Raymond Duvall, "Power in International Politics," *International organization* 59, no. 1 (2005): 39–75.
5. An exception would arise if there were demand specifically for human labor, such as a consumer preference for goods made "by hand."
6. Irving John Good, "Speculations concerning the First Ultraintelligent Machine," *Advances in Computers* 6, no. 99 (1965): 31–83.
7. William D. Nordhaus, *Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth,* no. w21547 (Cambridge, MA: National Bureau of Economic Research, 2015).
8. Robin Hanson, *The Age of Em: Work, Love, and Life When Robots Rule the World* (Oxford: Oxford University Press, 2016), ch. 16.
9. Nehal Bhuta et al., eds.. *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge: Cambridge University Press, 2016).
10. Ryan Calo, "Peeping HALs: Making Sense of Artificial Intelligence and Privacy," *European Journal of Legal Studies* 2, no. 3 (2010): 168.
11. FAT/ML, "Fairness, Accountability, and Transparency in Machine Learning," 2018, https://www.fatml.org/.
12. Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law and Technology* 29, no. 2 (2016): 353–400.
13. Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in Nick Bostrom and Milan M. Ćirković (eds.), *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 308–45; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, no. 4 (2015): 105–114..
14. For example, see House of Lords, Select Committee on Artificial Intelligence, *AI in the UK: Ready, Willing and Able?,* March 13, 2018, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.
15. We confine our analysis to desiderata that satisfy a basic universalizability criterion. For example, if there is some respect in which the special circumstances would give actor *A* stronger-than-usual reason to harm actor *B*, and give actor *B* stronger-than-usual reason to harm actor *A*, then there would in some sense be a general pattern that could be discerned and distilled into the policy recommendation "Put greater

emphasis on attacking each other." But in this generalized form, the policy change would not be desirable to anybody; so since it fails universalizability, we would not include it as a desideratum.

16. By "technological maturity" we mean the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved. Nick Bostrom, "Existential Risk Prevention as Global Priority," *Global Policy* 4, no. 1 (2013): 15–31.
17. Frank J. Tipler, "Extraterrestrial Intelligent Beings Do Not Exist," *Quarterly Journal of the Royal Astronomical Society* 21 (1980): 267–81; Stuart Armstrong and Anders Sandberg, "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox," *Acta Astronautica* 89 (2013): 1–13.
18. David Pearce, *The Hedonistic Imperative*, Hedweb, 1995, https://www.hedweb.com/hedab.htm; Nick Bostrom, "Transhumanist Values," *Journal of Philosophical Research* 30, Supplement (2005): 3–14; Nick Bostrom, "Letter from Utopia," *Studies in Ethics, Law, and Technology* 2, no. 1 (2008): 1–7.
19. Nick Beckstead, "On the Overwhelming Importance of Shaping the Far Future" (PhD diss., Rutgers University, 2013), chs. 4–5; Nick Bostrom, "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* 15, no. 3 (2003): 308–14.
20. Perhaps in digital form (Anders Sandberg and Nick Bostrom, "Whole Brain Emulation: A Roadmap," Technical Report 2008-3, Future of Humanity Institute, University of Oxford, 2008, http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf) or in biological form via advanced biotechnological or nanotechnological means (Kim Eric Drexler, *Engines of Creation: The Coming Era of Nanotechnology* [New York: Anchor Books, 1986], ch. 7; Robert A. Freitas, *Nanomedicine,* vol. 1: *Basic Capabilities* [Georgetown, TX: Landes Bioscience, 1999]). There is a sense in which it might already be possible for some currently existing individuals to reach astronomical lifespans, namely, by staying alive through ordinary means until an intelligence explosion or other technological breakthrough occurs. Also cryonics (Nick Bostrom, "The Transhumanist FAQ: v 2.1," World Transhumanist Association, 2003, http://www.nickbostrom.com/views/transhumanist.pdf; Ralf C. Merkle, "The Molecular Repair of the Brain," *Cryonics Magazine* 15 (1994): 16-31).
21. Actors with a very high discount for future life duration might, however, prefer to postpone superintelligence until they are at death's door, since the arrival of machine superintelligence might involve a momentarily elevated level of risk (cf. section 10.3.2.).
22. Bostrom, *Superintelligence*; Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010), 1036–40.
23. Stuart Armstrong, Nick Bostrom, and Carl Shulman, "Racing to the Precipice: A Model of Artificial Intelligence Development," *AI & Society* 31, no. 2 (2016): 201–6.
24. Nick Bostrom, "The Future of Human Evolution," in *Death and Anti-Death: Two Hundred Years after Kant, Fifty Years after Turing* (Palo Alto, CA: Ria University Press, 2004), 339–71; Scott Alexander, "Meditations on Moloch," *Slate Star Codex*, July 30, 2014, http://slatestarcodex.com/2014/07/30/meditations-on-moloch/.

25. For example, scholars and philanthropists should invest more in understanding global governance and possibilities for world government; policymakers should invest more in solving existing global coordination problems to provide practice and institutional experience for larger challenges; and fora for global governance should invest more in consideration of hypothetical coordination challenges.
26. Stabilization may involve centralizing control of the dangerous technology or instituting a monitoring regime that would enable the timely detection and interception of any move to deploy the technology for a destructive purpose; cf. Nick Bostrom, "The Vulnerable World Hypothesis." *Global Policy* 10, no. 4 (2019): 455–76.
27. An ideal alignment solution would enable control of both external and internal behavior (thus making it possible to avoid intrinsically undesirable types of computation without sacrificing much in terms of performance; cf. "mind crime" discussion).
28. A singleton is a world order which at the highest level has a single decision-making agency, with the ability to "prevent any threats (internal or external) to its own existence and supremacy" and to "exert effective control over major features of its domain (including taxation and territorial allocation)." Nick Bostrom, "What Is a Singleton," *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48.
29. To be clear, we do not claim that the desiderata we identify are the *only* distributional desiderata that should be taken into account. There may also be desiderata that derive their justification from some source other than the special circumstances obtaining in our superintelligent AI scenario. (There might also be some additional allocation-related desiderata that *could* have been derived from those special circumstances but that we have failed to include in this paper. We do not claim completeness.)
30. Note that care would have to be taken, when following the principle, not to implement it in a way that unduly inhibits socially desirable risk-taking, such as many forms of experimentation and innovation. Internalizing the negative externalities of such activities without also internalizing their positive externalities could produce worse incentives than if neither kind of externality were internalized.
31. Thomas Piketty, *Capital in the Twenty-First Century,* trans. A. Goldhammer (Cambridge, MA: Belknap Press, 2014), ch. 7.
32. Eric Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (Vancouver: W. W. Norton, 2014).
33. It could also reduce permutation after the transition to the machine intelligence era, if it is easier to bequeath capital to one's children (or to preserve it oneself while one is alive, which might be for a very long time with the advent of effective life-extension technology) than it is to bequeath or preserve talents and skills under historically more usual circumstances.
34. Actors could seek to preserve their position by continuously diversifying their holdings. However, there may be substantial constraints and frictions on achieving this, related to (1) constraints or costs to diversifying, (2) time lags in diversification, (3) willingness of some actors to gamble big. (1a) Some asset classes (e.g., stealth start-ups, private companies, stakes in some national economies) are not available for ownership or involve a costly search and investment process. (1b) Many actors face major diversification constraints. A firm or a country might be heavily

committed to one industry sector and be unable to effectively hedge its exposures or quickly reorient its activities to adapt to a rapidly changing competitive landscape. (2) Technological churn may move so quickly that investors do not have the opportunity to rebalance their portfolios "in time." By the time an appreciating new asset class is evident, one may have already lost out on much of its growth value. (3) Some actors will choose to make big bets on risky assets or technology, which if they win would reshuffle the ranking of wealth; even top-tier perfectly diversified actors could be deposed from their apex position by some upstart who scores a jackpot in the great reshuffle.

The distribution of military power is also in principle subject to reshuffling induced by accelerated technological churn, including in ways that are difficult to guard against by military diversification or using existing strength to bargain for a stable arrangement that locks in existing power hierarchies.

35. Bruce Bueno de Mesquita and Alistair Smith, *The Dictator's Handbook: Why Bad Behavior Is Almost Always Good Politics* (New York: Public Affairs, 2011); Michael C. Horowitz, "Who'll Want Artificially Intelligent Weapons? ISIS, Democracies, or Autocracies?," *Bulletin of the Atomic Scientist,* July 29, 2016, http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692.
36. We can distinguish two kinds of permutation. (1) Permutations where an individual's *expected* ex post wealth (or power, status, etc.) equals her ex ante wealth (power, status, etc.). Such a permutation is like a conventional lottery, where the more tickets you have, the more you can expect to win. Risk-averse individuals can try to hedge against such permutations by diversifying their holdings, but as noted in note 34, sufficiently drastic reshufflings can be hard to hedge against, especially in scenarios with large-scale violations of contracts and property rights. (2) Permutations where an individual's expected ex post wealth is unrelated to her ex ante wealth. Think of this as random role-switching: everybody's names are placed in a large urn, and each individual pulls out one ticket; she gives up what she had before and instead gets that other person's endowment. Setting aside the consequences of social disruption, this type of permutation would result in an expected gain for those who were initially poorly off, at the expense of incumbent elites. However, those who on nonselfish grounds favor redistribution to the poor typically want this to be done by reducing economic inequality rather by having a few of the poor swap places with the rich.
37. This is meant as an extension of the "veil of ignorance" thought experiment proposed by John Rawls: "The parties . . . do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations. . . . First of all, no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities." John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap Press, 1971), 137.
38. Such agreement is well established by, among other agreements, The United Nations, *Charter of the United Nations*, October 24, 1945, http://www.refworld.org/docid/3ae6b3930.html, and the *International Bill of Human Rights*, composed of the United

Nations General Assembly, *Universal Declaration of Human Rights*, 217 A (III), December 10, 1948, http://www.refworld.org/docid/3ae6b3712c.html, the United Nations General Assembly, International Covenant on Civil and Political Rights, Treaty Series, vol. 999, p. 171, December 16, 1966, http://www.refworld.org/docid/3ae6b3aa0.html, and the United Nations General Assembly, *International Covenant on Economic, Social and Cultural Rights*, Treaty Series, vol. 993, p. 3, December 16, 1966, http://www.refworld.org/docid/3ae6b36c0.html, which have been nearly universally ratified (though, significantly, not ratified by China or the United States). Further support for this can be found in the international legal principle of *jus cogens* (compelling law), which forms binding international legal norms from which no derogation is permitted. While the exact scope of jus cogens is debated, there is general consensus that it includes prohibitions against slavery, torture, and genocide, among other things. See Anne Lagerwall, "Jus Cogens," Oxford Bibliographies, May 29, 2015, http://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953–0124.xml. For more on the potential relationship between international human rights law and AI development as it relates to existential risk, see Silja Voeneky. "Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks." in Silja Voeneky, and Gerald L. Neuman (eds.), *Human Rights, Democracy, and Legitimacy in a World of Disorder*, 139–62. Cambridge: Cambridge University Press, 2018..

39. Hanson, *The Age of Em*, 189–94.
40. One technology area that one could expect to be brought to maturity within some years after the development of strongly superintelligent AI is advanced capabilities for space colonization, including the ability to emit von Neumann probes that are capable of traveling at some meaningful fraction of the speed of light over intergalactic distances and bootstrapping a technology base on a remote resource that is capable of producing and launching additional probes. R. Robert A. Freitas, "A Self-Reproducing Interstellar Probe," *Journal of the British Interplanetary Society* 33, no. 7 (1980): 251–64; Tipler, "Extraterrestrial Intelligent Beings Do Not Exist." Assuming the capability of creating such von Neumann probes, and that the observable universe is void of other intelligent civilizations (Anders Sandberg, Eric Drexler, and Toby Ord, "Dissolving the Fermi Paradox," *arXiv*, June 6, 2018, arXiv:1806.02404), then humanity's cosmic endowment would appear to include 10^{18} to 10^{20} reachable stars (Armstrong and Sandberg, "Eternity in Six Hours"). With the kind of astrophysical engineering technology that one would also expect to be available over the relevant timescales (Anders Sandberg, "*Grand Futures,*" unpublished manuscript, 2020), this resource base could suffice to create habitats for something like 10^{35} biological human lives (over the course of the remaining lifespan of the universe) or, alternatively, for a much larger number (in the vicinity of 10^{58} or more) of digitally implemented human minds (Bostrom, *Superintelligence*). Of course, most of this potential could be realized only over very long timescales, but for patient actors, the delays may not matter much.

    Note that a larger fraction of actors may be "patient" in the relevant sense after technological means for extreme life extension or suspended animation (e.g.,

facilitated by digital storage of human minds) are developed. Actors that anticipate that such capabilities will be developed shortly after the arrival of superintelligent AI may be patient—in the sense of not severely discounting temporally extremely remote economic benefits—in anticipation, since they might attach a nontrivial probability to themselves being around to consume some of those economic benefits after the long delay. Another important factor that could make extremely distant future outcomes decision-relevant to a wider set of actors is that a more stable social order or other reliable commitment techniques may become feasible, increasing the chance that near-term decisions could have predictable effects on what happens in the very long run.

41. The estimated 2017 world GDP was 80 trillion USD nominally (or 120 trillion USD when considering purchasing power parity, according to the World Bank. *Purchasing Power Parities and the Size of World Economies: Results from the 2017 International Comparison Program,* Washington, DC: World Bank, 2020). This is equivalent to a GDP per capita of 11,000 USD (nominal) or 17,000 USD (PPP). In order for a 40,000 USD guaranteed basic annual income to be achieved with 5% of world GDP at 2018 population levels (of 7.6 billion), world GDP would need to increase by a factor of 50 to 75, to 6 quadrillion (10^15) USD. While 5% may sound like a high philanthropic rate, it is actually half of the average of the current rate of the ten richest Americans. While the required increase in economic productivity may seem large, it requires just six doublings of the world economy. Over the past century, doublings in world GDP per person have occurred roughly every thirty-five years. Advanced machine intelligence would likely lead to a substantial increase in the growth rate of wealth per (human) person. The economist Robin Hanson has argued that after the arrival of human-level machine intelligence, in the form of human brain emulations, doublings could be expected to occur every year or even month (*The Age of Em*, 189–91).

    Note also that we are assuming here and elsewhere, perhaps unrealistically, that we are either not living in a computer simulation, or that we are but that it will continue to run for a considerable time after the development of machine superintelligence. Nick Bostrom, "Are We Living in a Computer Simulation?," *Philosophical Quarterly* 53, no. 211 (2003): 243–55. If we are in a simulation that terminates shortly after superintelligent AI is created, then the apparent cosmic endowment may be illusory; a different set of considerations then come into play, which are beyond the scope of this paper.

42. With sufficiently advanced technology, bioengineered meat substitutes should remove altogether the incompatibility between carnivorous consumer preferences and animal welfare. And with even more advanced technology, consumers might reengineer their taste buds to prefer ethical, healthy, sustainable plant foods, or (in the case of uploads or other digital minds) eat electricity and virtual steaks.

43. Here, epsilon-magnanimity might be seen as amounting to a weak form of practical value pluralism.

44. Nick Bostrom. *Moral Uncertainty—towards a Solution?* January 1, 2009. http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

45. An epsilon-magnanimous future could be achieved by ensuring that the future is shaped by many actors, representing many different values, each of whom is able to exert some nonnegligible degree of influence; or, alternatively, by ensuring that at least one extremely empowered actor is individually epsilon-magnanimous.
46. For example, it would appear both feasible and desirable under these circumstances to extend assistance to nonhuman animals, including wildlife, to mitigate their hardship, reduce suffering, and bring increased joy to all reachable sentient beings (Pearce, *The Hedonistic Imperative*).
47. In principle, these observations pertain also to biological minds insofar as they share the relevant properties. Conceivably, extremely advanced biotechnology might enable biological structures to approximate some of the attributes that would be readily available for digital implementations.
48. Bostrom, *Superintelligence.*
49. For examples of the mockery surrounding the earliest animal cruelty laws, see David R. Fisher, "Martin, Richard (1754–1834), of Dangan and Ballynahinch, Co. Galway and 16 Manchester Buildings, Mdx," in David R. Fisher (ed.), *The History of Parliament: The House of Commons 1820–1832* (Cambridge: Cambridge University Press, 2009), http://www.historyofparliamentonline.org/volume/1820-1832/member/martin-richard-1754-1834. For more on the changing norms regarding the treatment of animals, see Steven Pinker, *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes* (London: Penguin, 2011), chs. 3, 6.
50. But not all sophisticated minds need have such interests. We may assume that it is wrong to enslave or exploit human beings or other beings that are very similar to humans. But it may well be possible to design an AI with human-level intelligence (but differing from humans in other ways, such as in its motivational system) that would not have an interest in not being "enslaved" or "exploited." See also Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in William Ramsey and Keith Frankish (eds.), *The Cambridge Handbook of Artificial Intelligence* (Cambridge: Cambridge University Press, 2014), 316–34.
51. For evidence of the heritability of traits in modern society associated with larger family size, see Emmanuel Milot et al., "Evidence for Evolution in Response to Natural Selection in a Contemporary Human Population," *Proceedings of the National Academy of Sciences* 108, no. 41 (2011): 17040–45; Augustine Kong et al., "Selection against Variants in the Genome Associated with Educational Attainment," *Proceedings of the National Academy of Sciences* 114, no. 5 (2017): E727–E732. According to Jonathan P. Beauchamp, "Genetic Evidence for Natural Selection in Humans in the Contemporary United States," *Proceedings of the National Academy of Sciences* 113, no. 28 (2016): 7774: "In modern populations with low mortality, fitness can be reasonably approximated by [the number of children an individual ever gave birth to or fathered]."
52. The simple argument focuses on the possibility of economically unproductive beings, such as children, which is sufficient to establish the conclusion. But it is also possible to run into Malthusian problems when the minds generated are economically productive; see Hanson, *The Age of Em,* for a detailed examination of such a scenario.

Global coordination would be required to avoid the Malthusian outcome in the Hansonian model.

53. One example of a reproductive paradigm would be to require a would-be progenitor, prior to creating a new mind, to set aside a sufficient economic endowment to guarantee the new mind an adequate quality of life without further transfers. For as long as the world economy keeps growing, occasional "free" progeny could also be allowed, at a rate set to keep the population growth rate no higher than the economy growth rate.
54. A similar process can unfold with biological citizens, albeit over a longer timescale, if some group finds a way to keep its values stable while sustaining a high level of fertility.
55. Option (i) could take various forms. For instance, one could transition to a system in which voting rights are inherited. Some initial population would be endowed with voting rights (such as current people who have voting rights and their existing children upon coming of age). When one of these electors creates a new eligible being—whether a digital copy or surrogate, or a biological child—then the voting rights of the original are split between progenitor and progeny, so that the voting power of each "clan" remains constant. This would prevent fast-growing clans from effectively disenfranchising slower-growing populations and would remove the perverse incentive to multiply for the sake of gaining political influence.

    Robin Hanson has suggested the alternative of speed-weighted voting, which would grant more voting power to digital minds that run on faster computers (*The Age of Em*, 265). This may reduce the problem of voter inflation (by blocking one strategy for multiplying representation—running many slow, and therefore computationally cheap, copies). However, it would give extra influence to minds that are wealthy enough to afford fast implementation or that happen to serve in economic roles demanding fast implementation.
56. A fourth ground might be to ensure that decisions are perceived as legitimate.
57. These norms and principles may have gained traction because they helped with governance challenges within the sociotechnological millieus of previous decades and centuries.
58. Some of our discussion earlier in this paper offers additional examples of instances where existing norms would need to be rescinded or reconceived. The right to unlimited reproduction is hardly defensible in a context where Malthusian concerns loom large, such as for digital minds. Freedom of thought may similarly need to be circumscribed in the case of AI minds that have the ability merely by thinking about a suffering subject in great detail to create internally that mind in a state of suffering and thus engage in an act of mind crime. Punishment for criminal offenses: some of the current reasons for incarceration would cease to apply if, for instance, advanced AI made it possible to more effectively rehabilitate offenders or to let them back into society without endangering other citizens, or if the introduction of more effective crime prevention methods reduced the need to deter future crime. The meaning of a given sentence: even if a life sentence is sometimes a just punishment when the typical remaining lifespan is a few decades, it may not be just if AI-enabled medicine

makes it possible to greatly extend lifespan. Various dignity-based or religious sensitivities may require special protections and accommodations in the context of advanced AI. And AI research itself may need to be approached in a different manner than most basic research, where norms of curiosity-driven exploration, openness, and the celebration of intellectual achievement are often held up as the ultimate touchstones. For AI research, considerations about downstream applications and strategic impacts of research findings may need to be added to the criteria by which research contributions are evaluated.

59. For comment and discussion, we're grateful to Stuart Armstrong, Michael Barnett, Seth Baum, Dominic Becker, Nick Beckstead, Devi Borg, Miles Brundage, Paul Christiano, Jack Clark, Rebecca Crootof, Richard Danzig, Daniel Dewey, Eric Drexler, Sebastian Farquhar, Sophie Fischer, Ben Garfinkel, Katja Grace, Tom Grant, Hilary Greaves, Rose Hadshar, John Halstead, Robin Hanson, Verity Harding, Sean Legassick, Wendy Lin, Jelena Luketina, Matthijs Maas, Luke Muehlhauser, Toby Ord, Mahendra Prasad, Anders Sandberg, Carl Shulman, Andrew Snyder-Beattie, Nate Soares, Mojmir Stehlik, Jaan Tallinn, Alex Tymchenko, and several anonymous reviewers. This work has received funding, in part, from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 669751), and the Future of Life Institute.

References

Alexander, Scott. "Meditations on Moloch." *Slate Star Codex,* July 30, 2014. http://slatestarcodex.com/2014/07/30/meditations-on-moloch/.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI & Society* 31, no. 2 (2016): 201–6.

Armstrong, Stuart, and Anders Sandberg. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89 (2013): 1–13.

Barnett, Michael, and Raymond Duvall. "Power in International Politics." *International organization* 59, no. 1 (2005): 39–75.

Beauchamp, Jonathan P. "Genetic Evidence for Natural Selection in Humans in the Contemporary United States." *Proceedings of the National Academy of Sciences* 113, no. 28 (2016): 7774–79.

Beckstead, Nick "On the Overwhelming Importance of Shaping the Far Future." PhD diss., Rutgers University, 2013.

Bhuta, Nehal, Susanne Beck, Robin Geiβ, Hin-Yan Liu, and Claus Kreβ, eds. *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press, 2016.

Bostrom, Nick "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53, no. 211 (2003): 243–55.

Bostrom, Nick "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15, no. 3 (2003): 308–14.

Bostrom, Nick "Existential Risk Prevention as Global Priority." *Global Policy* 4, no. 1 (2013): 15–31.

Bostrom, Nick "The Future of Human Evolution." In *Death and Anti-Death: Two Hundred Years after Kant, Fifty Years after Turing*, 339–71. Palo Alto, CA: Ria University Press, 2004.

Bostrom, Nick "Letter from Utopia." *Studies in Ethics, Law, and Technology* 2, no. 1 (2008): 1–7.

Bostrom, Nick *Moral Uncertainty—towards a Solution?* January 1, 2009. http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

Bostrom, Nick *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, 2014.

Bostrom, Nick "The Transhumanist FAQ: v 2.1." World Transhumanist Association, 2003. http://www.nickbostrom.com/views/transhumanist.pdf.

Bostrom, Nick "Transhumanist Values." *Journal of Philosophical Research* 30, Supplement (2005): 3–14.

Bostrom, Nick "The Vulnerable World Hypothesis." *Global Policy* 10, no. 4 (2019): 455–76.

Bostrom, Nick "What Is a Singleton." *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54.

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In William Ramsey and Keith Frankish (eds.) *The Cambridge Handbook of Artificial Intelligence*, 316–34. Cambridge: Cambridge University Press, 2014.

Brynjolfsson, Eric, and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Vancouver: W. W. Norton, 2014.

Bueno de Mesquita, Bruce, and Alistair Smith. *The Dictator's Handbook: Why Bad Behavior Is Almost Always Good Politics*. New York: Public Affairs, 2011.

Calo, Ryan "Peeping HALs: Making Sense of Artificial Intelligence and Privacy." *European Journal of Legal Studies* 2, no. 3 (2010): 168.

Drexler, Kim Eric *Engines of Creation: The Coming Era of Nanotechnology*. New York: Anchor Books, 1986.

FAT/ML. "Fairness, Accountability, and Transparency in Machine Learning." 2018. https://www.fatml.org/.

Fisher, David R. "Martin, Richard (1754–1834), of Dangan and Ballynahinch, Co. Galway and 16 Manchester Buildings, Mdx." In David R. Fisher (ed.) *The History of Parliament: The House of Commons 1820–1832*. Cambridge: Cambridge University Press, 2009. http://www.historyofparliamentonline.org/volume/1820-1832/member/martin-richard-1754-1834.

Freitas, Robert A. *Nanomedicine.* Vol. 1: *Basic Capabilities.* Georgetown, TX: Landes Bioscience, 1999.

Freitas, Robert A. "A Self-Reproducing Interstellar Probe." *Journal of the British Interplanetary Society* 33, no. 7 (1980): 251–64.

Good, Irving John "Speculations concerning the First Ultraintelligent Machine." *Advances in Computers* 6, no. 99 (1965): 31–83.

Grace, Katja, Salvatier, John, Dafoe, Aallan, Zhang, Baobao, & Evans, Owain "When will AI exceed human performance? Evidence from AI experts." *Journal of Artificial Intelligence Research*, 62 (2018): 729-754.

Hale, Thomas, and David Held. *Handbook of Transnational Governance*. Cambridge, UK: Polity Press, 2011.

Hanson, Robin *The Age of Em: Work, Love, and Life When Robots Rule the World.* Oxford: Oxford University Press, 2016.

Horowitz, Michael C. "Who'll Want Artificially Intelligent Weapons? ISIS, Democracies, or Autocracies?" *Bulletin of the Atomic Scientist*, July 29, 2016. http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692.

House of Commons, Science and Technology Committee. *Robotics and Artificial Intelligence: Fifth Report of Session 2016–17*. October 12, 2016. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

House of Lords, Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing and Able?* March 13, 2018. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Kong, Augustine, Michael L. Frigge, Gudmar Thorleifsson, Hreinn Stefansson, Alexander I. Young, Florian Zink, Gudrun A. Jonsdottir, Aysu Okbay, Patrick Sulem, Gisli Masson, Daniel F. Gudbjartsson, Agnar Helgason, Gyda Bjornsdottir, Unnur Thorsteinsdottir, and Kari Stefansson "Selection against Variants in the Genome Associated with Educational Attainment." *Proceedings of the National Academy of Sciences* 114, no. 5 (2017): E727–E732.

Lagerwall, Anne. "Jus Cogens." Oxford Bibliographies, May 29, 2015. http://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953–0124.xml.

Merkle, Ralf C. "The Molecular Repair of the Brain." *Cryonics Magazine* 15 (1994): 16-31.

Milot, Emmanuel, Francine M. Mayer, Daniel H. Nussey, Mireille Boisvert, Fanie Pelletier, and Denis Réale. "Evidence for Evolution in Response to Natural Selection in a Contemporary Human Population." *Proceedings of the National Academy of Sciences* 108, no. 41 (2011): 17040–45.

Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, 553–71. Geneva: Springer International, 2016.

National Science and Technology Council. 2016. *Preparing for the Future of Artificial Intelligence*. Washington, DC: Office of Science and Technology Policy, 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Nordhaus, William D. 2015. *Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth*. No. w21547. Cambridge, MA: National Bureau of Economic Research, 2015.

Pearce, David 1995. *The Hedonistic Imperative*. Hedweb, 1995. https://www.hedweb.com/hedab.htm.

Piketty, Thomas 2014. *Capital in the Twenty-First Century.* Translated by A. Goldhammer. Cambridge, MA: Belknap Press, 2014.

Pinker, Steven 2011. *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes.* London: Penguin, 2011.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press, 1971.

Roff, Heather M. "The Strategic Robot Problem: Lethal Autonomous Weapons in War." *Journal of Military Ethics* 13, no. 3 (2014): 211–27.

Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36, no. 4 (2015): 105–114.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

Sandberg, Anders "*Grand Futures*." Unpublished manuscript, 2020

Sandberg, Anders, and Nick Bostrom. "Whole Brain Emulation: A Roadmap." Technical Report 2008-3. Future of Humanity Institute, University of Oxford, 2008. http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf.

Sandberg, Anders, Eric Drexler, and Toby Ord. "Dissolving the Fermi Paradox." *arXiv,* June 6, 2018. arXiv:1806.02404.

Scherer, Matthew U. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law and Technology* 29, no. 2 (2016): 353–400.

Tipler, Frank J. "Extraterrestrial Intelligent Beings Do Not Exist." *Quarterly Journal of the Royal Astronomical Society* 21 (1980): 267–81.

United Nations. *Charter of the United Nations.* October 24, 1945. http://www.refworld.org/docid/3ae6b3930.html.

United Nations General Assembly. *International Covenant on Civil and Political Rights*. Treaty Series, vol. 999, p. 171, December 16, 1966. http://www.refworld.org/docid/3ae6b3aa0.html.

United Nations General Assembly. *International Covenant on Economic, Social and Cultural Rights*. Treaty Series, vol. 993, p. 3, December 16, 1966. http://www.refworld.org/docid/3ae6b36c0.html.

United Nations General Assembly. *Universal Declaration of Human Rights*. 217 A (III). December 10, 1948. http://www.refworld.org/docid/3ae6b3712c.html.

Voeneky, Silja. "Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks." in Silja Voeneky, and Gerald L. Neuman (eds.), *Human Rights, Democracy, and Legitimacy in a World of Disorder*, 139–62. Cambridge: Cambridge University Press, 2018..

West, Darrell M., and John R. Allen. *How Artificial Intelligence Is Transforming the World.* Brookings Institute, April 24, 2018. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.

World Bank. *Purchasing Power Parities and the Size of World Economies: Results from the 2017 International Comparison Program,* Washington, DC: World Bank, 2020.

Yudkowsky, Eliezer "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Nick Bostrom and Milan M. Ćirković (eds.), *Global Catastrophic Risks,* 308–45. Oxford: Oxford University Press, 2008.

11
Artificial Intelligence
A Binary Approach

Stuart Russell

11.1. Introduction

Artificial intelligence (AI) has as its aim the creation of intelligent machines. An entity is intelligent, roughly speaking, if it chooses actions that are expected to achieve its objectives, given what it has perceived. Applying this definition to machines, one can deduce that AI aims to create machines that choose actions that are expected to achieve their objectives, given what they have perceived.

Now, what are these objectives? To be sure, they are—up to now, at least—objectives that we put into them; nonetheless they are objectives that operate exactly as if they were the machines' own and about which the machines are completely certain.

In 1960, after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator, Norbert Wiener gave a clear warning: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire."[1] In my view, this is the source of the existential risk from superintelligent AI cited in recent years by such observers as Elon Musk,[2] Bill Gates,[3] Stephen Hawking,[4] and Nick Bostrom.[5] There is very little chance that we humans can specify our objectives completely and correctly, in such a way that the pursuit of those objectives by more capable machines is guaranteed to result in beneficial outcomes for humans.

The mistake comes from transferring a perfectly reasonable definition of intelligence from humans to machines. The definition is reasonable for humans because we are entitled to pursue our own objectives—indeed, whose would we pursue, if not our own? The definition is *unary*, in the sense that it applies to an entity by itself. Machines, on the other hand, are not entitled to pursue their own objectives. A sensible definition of AI would have machines pursuing *our* objectives. Thus we have a *binary* definition: entity A chooses actions that are expected to achieve the objectives of entity B, given what entity A has perceived. In the unlikely event that we (entity B) can specify the objectives completely and correctly

Stuart Russell, *Artificial Intelligence* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0012

and insert them into the machine (entity A), then we can recover the original, unary definition. If not, then the machine will necessarily be *uncertain* as to our objectives, while being obliged to pursue them on our behalf. This uncertainty, and the resulting coupling between machines and humans that it entails, is crucial to building AI systems of arbitrary intelligence that are provably beneficial to humans. We must therefore reconstruct the foundations of AI along binary rather than unary lines.

11.2. Artificial Intelligence

The goal of AI research has been to understand the principles underlying intelligent behavior and to build those principles into machines that can then exhibit such behavior. In the 1960s and 1970s the prevailing theoretical definition of intelligence was the capacity for logical reasoning, including the ability to derive plans of action guaranteed to achieve a specified goal. More recently a consensus has emerged in AI around the idea of a rational agent that perceives, and acts in order to maximize, its expected utility. Subfields such as logical planning, robotics, and natural-language understanding are special cases of the general paradigm. AI has incorporated probability theory to handle uncertainty, utility theory to define objectives, and statistical learning to allow machines to adapt to new circumstances. These developments have created strong connections to other disciplines that build on similar concepts, including control theory, economics, operations research, and statistics.

In both the logical-planning and rational-agent views of AI, the machine's objective—whether in the form of a goal, a utility function, or a reward function (as in reinforcement learning)—is specified exogenously. In Wiener's words, this is "the purpose put into the machine." Indeed it has been one of the tenets of the field that AI systems should be general-purpose—that is, capable of accepting a purpose as input and then achieving it—rather than special-purpose, with their goal implicit in their design. For example, a self-driving car should accept a destination as input instead of having one fixed destination. However, some aspects of the car's "driving purpose" are fixed, such as that it shouldn't hit pedestrians. This is built directly into the car's steering algorithms rather than being explicit: no self-driving car in existence today "knows" that pedestrians prefer not to be run over.

Putting a purpose into a machine that optimizes its behavior according to clearly defined algorithms seems an admirable approach to ensuring that the machine's "conduct will be carried out on principles acceptable to us." But, as Wiener warns, we need to put in the right purpose. We might call this the King Midas problem: Midas got exactly what he asked for—namely, that everything

he touched would turn to gold—but, too late, he discovered the drawbacks of drinking liquid gold and eating solid gold. The technical term for putting in the right purpose is *value alignment*. When it fails, we may inadvertently imbue machines with objectives counter to our own. Tasked with finding a cure for cancer as fast as possible, an AI system might elect to use the entire human population as guinea pigs for its experiments. Asked to de-acidify the oceans, it might use up all the oxygen in the atmosphere as a side effect. This is a common characteristic of systems that optimize: variables not included in the objective may be set to extreme values to help optimize that objective.

Unfortunately, neither AI nor other disciplines (economics, statistics, control theory, operations research) built around the optimization of objectives have much to say about how to identify the purposes "we really desire." Instead they assume that objectives are simply implanted into the machine. AI research, in its present form, studies the ability to achieve objectives, not the design of those objectives.

Steve Omohundro has pointed to a further difficulty, observing that intelligent entities must act to preserve their own existence.[6] This tendency has nothing to do with a self-preservation instinct or any other biological notion; it's just that an entity cannot achieve its objectives if it's dead. According to Omohundro's argument, a superintelligent machine that has an off-switch—which some, including Alan Turing himself,[7] have seen as our potential salvation—will take steps to disable the switch in some way. Thus we may face the prospect of superintelligent machines—their actions by definition unpredictable and their imperfectly specified objectives conflicting with our own—whose motivation to preserve their existence in order to achieve those objectives may be insuperable.

11.3. 1,001 Reasons to Pay No Attention

Objections have been raised to these arguments, primarily by researchers within the AI community. The objections reflect a natural defensive reaction, coupled perhaps with a lack of imagination about what a superintelligent machine could do. None holds water on closer examination. Here are some of the more common ones:

- *Don't worry, we can just switch it off.*[8] This is often the first thing that pops into a layperson's head when considering risks from superintelligent AI—as if a superintelligent entity would never think of that. It's rather like saying that the risk of losing to Deep Blue or AlphaGo is negligible; all one has to do is make the right moves.

- *Human-level or superhuman AI is impossible*.[9] This is an unusual claim for AI researchers to make, given that, from Turing onward, they have been fending off such claims from philosophers and mathematicians. The claim, which is backed by no evidence, appears to concede that if superintelligent AI *were* possible, it *would* be a significant risk. It's as if a bus driver, with all of humanity as his passengers, said, "Yes, I'm driving toward a cliff—in fact, I'm pressing the pedal to the metal. But trust me, we'll run out of gas before we get there." The claim also represents a foolhardy bet against human ingenuity. We've made such bets before and lost. On September 11, 1933, the renowned physicist Ernest Rutherford stated, with utter confidence, "Anyone who expects a source of power from the transformation of these atoms is talking moonshine." On September 12, 1933, Leo Szilard invented the neutron-induced nuclear chain reaction. A few years later he demonstrated such a reaction in his laboratory at Columbia University. As he recalled in a memoir, "We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief."
- *It's too soon to worry about it*. The right time to worry about a potentially serious problem for humanity depends not just on when the problem will occur but also on how much time is needed to devise and implement a solution that avoids the risk. For example, if we were to detect a large asteroid predicted to collide with the Earth in 2067, would we say, "It's too soon to worry"? And if we consider the global catastrophic risks from climate change predicted to occur later in this century, is it too soon to take action to prevent them? On the contrary, it may be too late. The relevant timescale for human-level AI is less predictable, but, like nuclear fission, it might arrive considerably sooner than expected. One variation on this argument is Andrew Ng's statement that it's "like worrying about overpopulation on Mars." This appeals to a convenient analogy: Not only is the risk easily managed and far in the future, but also it's extremely unlikely that we'd even try to move billions of humans to Mars in the first place. The analogy is a false one, however. We're *already* devoting huge scientific and technical resources to creating ever more capable AI systems. A more apt analogy would be a plan to move the human race to Mars with no consideration for what we might breathe, drink, or eat once we arrived.
- *Human-level AI isn't really imminent, in any case.* The AI100 report, for example, assures us, "Contrary to the more fantastic predictions for AI in the popular press, the Study Panel found no cause for concern that AI is an imminent threat to humankind."[10] This argument simply misstates the reasons for concern, which are not predicated on imminence. In his 2014 book, *Superintelligence,* Nick Bostrom, for one, writes, "It is no part of the argument in this book that we are on the threshold of a big breakthrough in artificial intelligence, or that we can predict with any precision when such a development might occur."[11]

- *Any machine intelligent enough to cause trouble will be intelligent enough to have appropriate and altruistic objectives.*[12] (Often the argument adds the premise that people of greater intelligence tend to have more altruistic objectives, a view that may be related to the self-conception of those making the argument.) This argument is related to Hume's is-ought problem and G. E. Moore's naturalistic fallacy, suggesting that somehow the machine, as a result of its intelligence, will simply *perceive* what is right given its experience of the world. This is implausible; for example, one cannot perceive, in the design of a chessboard and chess pieces, the goal of checkmate; the same chessboard and pieces can be used for suicide chess, or indeed many other games still to be invented. Or consider another example: Nick Bostrom's thought experiment in which humans are driven to extinction by a putative robot that turns the planet into a sea of paper clips. We humans see this outcome as tragic, whereas the iron-eating bacterium *Thiobacillus ferrooxidans* is thrilled. Who's to say the bacterium is wrong? The fact that a machine has been given a fixed objective by humans doesn't mean that it will automatically recognize the importance to humans of things that aren't part of the objective. Maximizing the objective may well cause problems for humans, but, by definition, the machine will not recognize those problems as problematic.
- *Intelligence is multidimensional, "so 'smarter than humans' is a meaningless concept."*[13] It is a staple of modern psychology that IQ does not do justice to the full range of cognitive skills that humans possess to varying degrees. IQ is indeed a crude measure of human intelligence, but it is utterly meaningless for current AI systems, because their capabilities across different areas are uncorrelated. How do we compare the IQ of Google's search engine, which cannot play chess, with that of Deep Blue, which cannot answer search queries?

None of this supports the argument that because intelligence is multifaceted, we can ignore the risk from superintelligent machines. If "smarter than humans" is a meaningless concept, then "smarter than gorillas" is also meaningless, and gorillas therefore have nothing to fear from humans; clearly, that argument doesn't hold water. Not only is it logically possible for one entity to be more capable than another across all the relevant dimensions of intelligence, but it is also possible for one species to represent an existential threat to another even if the former lacks an appreciation for music and literature.

11.4. Solutions

Can we tackle Wiener's warning head-on? Can we design AI systems whose purposes don't conflict with ours, so that we're sure to be happy with how they behave? On the face of it, this seems hopeless, because it will doubtless prove

infeasible to write down our purposes correctly or imagine all the counterintuitive ways a superintelligent entity might fulfill them.

If we treat superintelligent AI systems as if they were black boxes from outer space, then indeed we have no hope. Instead the approach we seem obliged to take, if we are to have any confidence in the outcome, is to define some formal problem *F* and design AI systems to be *F*-solvers, such that no matter how perfectly a system solves *F*, we're guaranteed to be happy with the solution. If we can work out an appropriate *F* that has this property, we will be able to create *provably beneficial* AI.

There is, I believe, an approach that may work. Humans can reasonably be described as having (mostly implicit and partial) preferences over their future lives; that is, given enough time and unlimited visual aids, a human could express a preference (or indifference) when offered a choice between two future lives laid out before him or her in all their aspects. (This idealization ignores the possibility that our minds are composed of subsystems with effectively incompatible preferences; if true, that would limit a machine's ability to optimally satisfy our preferences, but it doesn't seem to prevent us from designing machines that avoid catastrophic outcomes.) The formal problem *F* to be solved by the machine in this case is to maximize human future-life preferences subject to its initial uncertainty as to what they are. Furthermore, although the future-life preferences are hidden variables, they're grounded in a voluminous source of evidence, namely, all of the human choices ever made. This formulation sidesteps Wiener's problem: the machine may learn more about human preferences as it goes along, of course, but it will never achieve complete certainty.

As noted in the introduction, this involves a shift from a unary view of AI to a binary one. The classical view, in which a fixed objective is given to the machine, is illustrated qualitatively in Figure 11.1. Once the machine has a fixed objective,

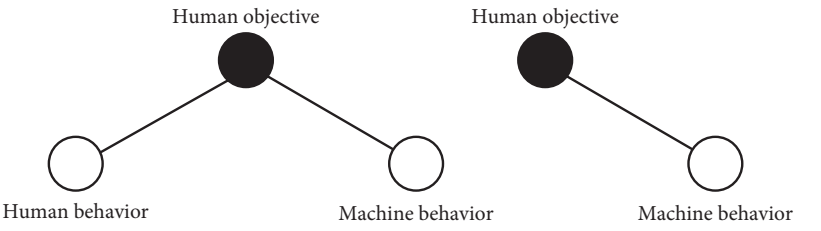


Figure 11.1 (a) The classical AI situation in which the human objective is considered fixed and known by the machine, depicted as a notional graphical model. Given the objective, the machine's behavior is (roughly speaking) independent of any subsequent human behavior, as depicted in (b). This unary view of AI is tenable only if the human objective can be be completely and correctly stated.

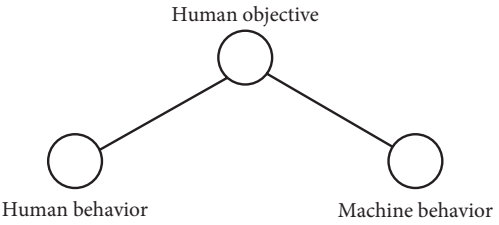


Figure 11.2 When the human objective is unobserved, machine behavior is no longer independent of human behavior, because the latter provides more information about the human objective.

it will act to optimize the achievement of the objective; its behavior is effectively independent of the human's behavior.[14]

This basic idea is made more precise in the framework of cooperative inverse reinforcement learning, or CIRL.[15] A CIRL problem involves two agents, one human and the other a robot. Because there are two agents, the problem is what economists call a *game*. It is in fact a game of *partial information*, because, while the human knows the reward function, the robot does not—even though the robot's job is to maximize it. It involves a form of inverse reinforcement learning because the robot can learn more about human preferences from the observation of human behavior—a process that is the dual of reinforcement learning, wherein behavior is learned from rewards and punishments.

A simple example will help to illustrate the basic idea. Suppose that Harriet, the human, likes to collect paper clips and staples, and her reward function depends on how many of each she has. More precisely, if she has p paper clips and s staples, her degree of happiness is $\theta p + (1\text{-}\theta)s$, where θ is essentially an exchange rate between paper clips and staples. If θ is 1, she likes only paper clips; if θ is 0, she likes only staples; if θ is 0.5, she is indifferent between them; and so on. It is the job of Robby, the robot, to produce the paper clips and staples. The point of the game is that Robby wants to make Harriet happy, but he does not know the value of θ, so he is not sure how many of each to produce.

Let's construct a specific scenario to see how the game works. Let the true value of θ be 0.49; that is, Harriet has a slight preference for staples over paper clips. And let's assume that Robby has a uniform prior belief about θ; that is, he believes θ is equally likely to be any value between 0 and 1. Harriet now gets to do a small demonstration, producing two paper clips, two staples, or one of each; after that, the robot can produce ninety paper clips, ninety staples, or fifty of each. Now one might think that Harriet, who prefers staples to paper clips, should produce two staples; of the three choices, this is the one she prefers. But in that case, Robby's rational response would be to produce ninety staples (with total value to Harriet

of 45.9), which is a worse outcome for Harriet than fifty of each (total value 50.0). The optimal solution of this particular game is that Harriet makes one of each if θ lies between 0.446 and 0.554, then Robby makes fifty of each. Thus the way the game is defined encourages Harriet to "teach" Robby—as long as she knows Robby is watching carefully.

Within the CIRL framework, one can formulate and solve the off-switch problem—that is, the problem of how to prevent a robot from disabling its off-switch. (Turing may rest easier.) A robot that is uncertain about human preferences actually benefits from being switched off, because it understands that the human will press the off-switch to prevent the robot from doing something counter to those preferences. The robot wants to avoid doing anything counter to human preferences, even if it doesn't know what those are. Thus the robot is incentivized to preserve the off-switch, and this incentive derives directly from its uncertainty about human preferences.[16]

The off-switch example suggests some templates for controllable-agent designs and provides at least one case of a provably beneficial system in the sense introduced earlier. The overall approach resembles mechanism-design problems in economics, wherein one incentivizes other agents to behave in ways beneficial to the designer. The key difference here is that we are building one of the agents in order to benefit the other.

11.5. Reasons for Optimism

There are some reasons to think this approach may work in practice. First, there is abundant written and filmed information about humans doing things (and other humans reacting). More or less every book ever written contains evidence on this topic. Even the oldest clay tablets, tediously recording the exchange of N sheep for M oxen, give information about human preferences between sheep and oxen. Technology to build models of human preferences from this storehouse will presumably be available long before superintelligent AI systems are created.

Second, there are strong near-term economic incentives for robots to understand human preferences, which also come into play well before the arrival of superintelligence. Already computer systems record one's preferences for an aisle seat or a vegetarian meal. More sophisticated personal assistants will need to understand their user's preferences for cost, luxury, and convenient location when booking hotels, and how these preferences depend on the nature and schedule of the user's planned activities. Managing a busy person's calendar and screening calls and emails requires an even more sophisticated understanding of the user's life, as does the management of an entire household when entrusted to a domestic robot. For all such roles, trust is essential but easily lost if the machine

reveals itself to lack a basic understanding of human preferences. If one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic-robot industry will be out of business.

11.6. Obstacles

There are obvious difficulties, however, with an approach that expects machines to learn underlying preferences from observing human behavior. The first is that humans are irrational, in the sense that our actions do not reflect our preferences. This irrationality arises in part from our computational limitations relative to the complexity of the decision problems we face. For example, if two humans are playing chess and one of them loses, it's because the loser (and possibly the winner too) made a mistake—a move that led inevitably to losing. A machine observing that move and assuming perfect rationality on the part of the human might well conclude that the human *preferred* to lose. Thus, to avoid reaching such conclusions, the machine must take into account the *actual* cognitive mechanisms of humans.

As yet we do not know enough about human cognitive mechanisms to invert real human behavior to get at the underlying preferences. One thing that seems intuitively clear, however, is that one of our principal methods for coping with the complexity of the world is to organize our behavior *hierarchically*. That is, we make (defeasible) commitments to higher-level goals such as "Write an essay on a binary approach to AI"; then, rather than considering all possible sequences of words, from "aardvark aardvark aardvark . . ." to "zyzzyva zyzzyva zyzzyva . . . ," as a chess program would do, we choose among subtasks such as "Write the introduction" and "Read Wiener's book." Eventually we get down to the choice of words, and then typing each word involves a sequence of keystrokes, each of which is in turn a sequence of motor control commands to the muscles of the arms and hands. At any given point, then, a human is embedded at various particular levels of multiple deep and complex hierarchies of partially overlapping activities and subgoals. This means that for the machine to understand human actions, it probably needs to understand a good deal about what these hierarchies are and how we use them to navigate the real world.

Machines might try to discover more about human cognitive mechanisms by an inductive learning approach. Suppose that in some given situation s Harriet's action a depends on her preferences θ according to mechanism h; that is, $a = h(\theta, s)$. (Here, θ represents not a single parameter such as the exchange rate between staples and paper clips, but Harriet's preferences over future lives, which could be a structure of arbitrary complexity.) By observing many examples of s and a, is

it possible eventually to recover h and θ? At first glance, the answer seems to be no. For example, one cannot distinguish between the following hypotheses about how Harriet plays chess:

1. h maximizes the satisfaction of preferences, and θ is the desire to win games.
2. h minimizes the satisfaction of preferences, and θ is the desire to lose games.

From the outside, Harriet plays perfect chess under either hypothesis.[17] If we are merely concerned with *predicting* her next move, it doesn't matter which formulation we choose. On the other hand, for a machine whose goal is to *help Harriet realize her preferences*, it really does matter! The machine needs to know which explanation holds. From this viewpoint, something is seriously wrong with the second explanation of behavior. If Harriet's cognitive mechanism h were really trying to minimize the satisfaction of preferences θ, it wouldn't make sense to call θ her preferences. It is, then, simply a mistake to suppose that h and θ are separately and independently defined. I have already argued that the assumption of perfect rationality—that is, h is maximization—is too strong; yet for it to make sense to say that Harriet has preferences, h will have to satisfy (or nearly satisfy) *some* basic properties associated with rationality. These might include choosing correctly according to preferences in situations that are computationally trivial—for example, choosing between vanilla and bubble-gum ice cream at the beach.[18]

Further difficulties arise if the machine succeeds in identifying Harriet's preferences but finds them to be inconsistent. For example, suppose she prefers vanilla to bubble-gum and bubble-gum to pistachio, but prefers pistachio to vanilla. In that case her preferences violate the axiom of transitivity and there is no way to maximally satisfy her preferences. (That is, whatever ice cream the machine gives her, there is always another that she would prefer.) In such cases, the machine could attempt to satisfy Harriet's preferences *up to inconsistency*; for example, if Harriet strictly prefers all three of those flavors to licorice, then it should avoid giving her licorice ice cream.

Of course, the inconsistency in Harriet's preferences could be of a far more radical nature. Many theories of cognition, such as Minsky's Society of Mind, posit multiple cognitive subsystems that, in essence, have their own preference structures and compete for control—and these seem to be manifested in addictive and self-destructive behaviors, among others. Such inconsistencies place limits on the extent to which the idea of machines helping humans even makes sense.

Also difficult, from a philosophical viewpoint, is the apparent *plasticity* of human preferences—the fact that they seem to change over time as the result of experiences. It is hard to explain how such changes can be made *rationally*,

because they make one's future self less likely to satisfy one's present preferences about the future. Yet plasticity seems fundamentally important to the entire enterprise, because newborn infants certainly lack the rich, nuanced, culturally informed preference structures of adults. Indeed it seems likely that our preferences are at least partially formed by a process resembling inverse reinforcement learning, whereby we absorb preferences that explain the behavior of those around us. Such a process would tend to give cultures some degree of autonomy from the otherwise homogenizing effects of our dopamine-based reward system.

Plasticity also raises the obvious question of which Harriet the machine should try to help: Harriet_{2020}, Harriet_{2025}, or some time-averaged Harriet? Plasticity is also problematic because of the possibility that by subtly influencing Harriet's environment, the machine may gradually mold her preferences in directions that make them easier to satisfy, much as certain political forces have been said to do with voters in recent decades.

I am often asked, "Whose values should we align AI with?" (The question is usually posed in more accusatory language, as if my secret, Silicon Valley–hatched plan is to align all the world's AI systems with my own white, male, Western, cisgender, Episcopalian values.) Of course, this is simply a misunderstanding. The kind of AI system proposed here is not "aligned" with *any* values, unless you count the basic principle of helping humans realize their preferences. For *each* of the billions of humans on Earth, the machine should be able to predict, to the extent that its information allows, which life that person would prefer.

Now, practical and social constraints will prevent all preferences from being maximally satisfied simultaneously, which means that machines must mediate among conflicting preferences—something that philosophers and social scientists have struggled with for millennia. At one extreme, each machine could pay attention only to the preferences of its owner, subject to legal constraints on its actions. This seems undesirable, as it would have a machine belonging to a misanthrope refuse to aid a severely injured pedestrian so that it can bring the newspaper home more quickly. Moreover we might find ourselves needing many more laws as machines satisfy their owners' preferences in ways that are very annoying to others even if not strictly illegal. At the other extreme, if machines consider equally the preferences of all humans, they will focus all their energies on the least fortunate and completely ignore their owners—a state of affairs not conducive to investment in AI. Presumably some middle ground can be found, perhaps combining a degree of obligation to the machine's owner with public subsidies that support contributions to the greater good.

Another common question is "What if machines learn from evil people?" Here, there is a real issue. It is *not* that machines will learn to copy evil actions. The machine's actions need not resemble in any way the actions of those from

whom it learns about human preferences, because it is trying to satisfy *their* preferences; it is not *adopting* those preferences as its own and acting to satisfy them. For example, suppose that a corrupt passport official in a developing country insists on a bribe for every transaction, so that he can afford to pay for his children's education. A machine observing this will not learn to take bribes itself: it has no need of money and understands (and wishes to avoid) the toll imposed on others by the taking of bribes. The machine will instead find other, socially beneficial ways to help send the children to school. Similarly, a machine observing humans killing each other in war will not learn that killing is good: obviously, those on the receiving end very much prefer not to be dead.

The difficult issue that remains is this: What should machines learn from humans who enjoy the suffering of others? In such cases, any simple aggregation scheme for preferences (such as adding utilities) would lead to some reduction in the utilities of others in order to satisfy, at least partially, these perverse preferences. It seems reasonable to require that machines simply ignore positive weights in the preferences of some for the suffering of others.

11.7. Looking Further Ahead

If we assume, for the sake of argument, that all of these obstacles can be overcome, as well as all of the obstacles to the development of truly capable AI systems, are we home free? Would provably beneficial, superintelligent AI usher in a golden age for humanity? Not necessarily. There remains the issue of adoption: how can we obtain broad agreement on suitable design principles, and how can we ensure that only suitably designed AI systems are deployed?

On the question of obtaining agreement at the policy level, it is necessary first to generate consensus within the research community on the basic ideas of—and design templates for—provably beneficial AI, so that policymakers have some concrete guidance on what sorts of regulations might make sense. The economic incentives noted earlier are of the kind that would tend to support the installation of rigorous standards, because failures would be damaging to entire industries, not just the perpetrator and victim. We already see this in miniature with the imposition of machine-checkable software standards for cell-phone applications.

On the question of enforcing policies for AI software design, I am less sanguine. If Dr. Evil wants to take over the world, he or she might remove the safety catch, so to speak, and deploy an AI system that ends up destroying the world instead. This problem is a hugely magnified version of the problem we currently face with malware. Our track record in solving the latter problem does not provide grounds for optimism concerning the former. In Samuel Butler's *Erewhon* and in Frank Herbert's *Dune*, the solution is to ban all intelligent machines, as

a matter of both law and cultural imperative. Perhaps if we find institutional solutions to the malware problem, we will be able to devise some less drastic approach for AI.

The problem of misuse is not limited to evil masterminds. One possible future for humanity in the age of superintelligent AI is that of a race of lotus eaters, progressively enfeebled as machines take over the management of our entire civilization. We may say, now, that such a future is undesirable; the machines may agree with us and volunteer to stand back, requiring humanity to exert itself and maintain its vigor. But exertion is tiring, and we may, in our usual myopic way, design AI systems that are not *quite* so concerned about the long-term vigor of humanity and just a *little* more helpful than they would otherwise wish to be. Unfortunately, this process continues in a direction that is hard to resist.

11.8. Summary

Finding a solution to the AI control problem is an important task; it may be, in Bostrom's words, "the essential task of our age." It involves building systems that are far more powerful than we are while still guaranteeing that those systems will remain powerless, forever.

Up to now, AI research has focused on systems that are better at making decisions, but this is not the same as making better decisions. No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if it fails to understand human preferences.

This problem requires a change in the definition of AI itself, from a field concerned with a unary notion of intelligence as the optimization of a given objective to a field concerned with a binary notion of machines that are provably beneficial for humans. Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship to it.

Notes

1. Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131 (1960): 1355–58.
2. Greg Kumparak, "Elon Musk Compares Building Artificial Intelligence to 'Summoning the Demon,'" *TechCrunch*, October 26, 2014.
3. Bill Gates, Reddit AMA, January 28, 2015: "I am in the camp that is concerned about superintelligence. . . . I agree with Elon Musk and some others on this and don't understand why some people are not concerned."

4. Hannah Osborne, "Stephen Hawking AI Warning: Artificial Intelligence Could Destroy Civilization." *Newsweek*, November 7, 2017.
5. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
6. Steve Omohundro, "The Basic AI Drives," in *Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin (Amsterdam; IOS Press, 2008), pp. 483–492.
7. Alan Turing, "Can Digital Computers Think?," *BBC Third Programme*, May 15, 1951, http://www.turingarchive.org/browse.php/B/5.
8. AI researcher Jeff Hawkins, for example, writes, "Some intelligent machines will be virtual, meaning they will exist and act solely within computer networks. . . . It is always possible to turn off a computer network, even if painful." Jeff Hawkins, "The Terminator Is Not Coming. The Future Will Thank Us." *Vox*, March 2, 2015.
9. The well-known AI100 report includes the following: "Unlike in the movies, there is no race of superhuman robots on the horizon or probably even possible." Peter Stone et al., "Artificial intelligence and life in 2030," One Hundred Year Study on Artificial Intelligence, report of the 2015 Study Panel, 2016, p.6.
10. Stone et al., *op. cit.*, p.4.
11. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), p.v.
12. Rodney Brooks, for example, asserts that it's impossible for a program to be "smart enough that it would be able to invent ways to subvert human society to achieve goals set for it by humans, without understanding the ways in which it was causing problems for those same humans." Rodney Brooks, "The Seven Deadly Sins of Predicting the Future of AI," *Robots, AI, and Other Stuff*, September 7, 2017, http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/.
13. Kevin Kelly, "The Myth of a Superhuman AI," *Wired*, April 25, 2017, https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.
14. The independence is not strict because the human's behavior can provide information about the state of the world. Thus a passenger in an automated taxi could tell the taxi that snipers have been reported on the road it intends to take, picking off passengers for fun, but this might affect the taxi's behavior only if it already knows that death by gunfire is undesirable for humans.
15. Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "Cooperative Inverse Reinforcement Learning," in *Advances in Neural Information Processing Systems 29*, edited by Daniel Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Cambridge, MA: MIT Press, 2017), pp. 3909–3917.
16. See Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "The Off-Switch Game," *arXiv*, last revised June 16, 2017, arXiv:1611.08219v3, https://arxiv.org/pdf/1611.08219.pdf.
17. Of course, the Harriet who prefers to lose might grumble when she keeps winning, thereby giving a clue as to which Harriet she is. One response to this is that grumbling is just more behavior, equally subject to multiple interpretations. This is not to say

that there is no fact of the matter as to whether Harriet is pleased or displeased with the outcome.

18. See, for example, Christopher Cherniak, *Minimal Rationality* (Cambridge, MA: MIT Press, 1986).

References

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

Brooks, Rodney. "The Seven Deadly Sins of Predicting the Future of AI." *Robots, AI, and Other Stuff*, September 7, 2017. http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/.

Cherniak, Christopher. *Minimal Rationality.* Cambridge, MA: MIT Press, 1986.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "Cooperative Inverse Reinforcement Learning." In *Advances in Neural Information Processing Systems 29*, edited by Daniel Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Cambridge, MA: MIT Press, 2017), pp. 3909–3917.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "The Off-Switch Game." *arXiv*, last revised June 16, 2017. arXiv:1611.08219v3. https://arxiv.org/pdf/1611.08219.pdf.

Kelly, Kevin. "The Myth of a Superhuman AI." *Wired*, April 25, 2017. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.

Kumparak, Greg. "Elon Musk Compares Building Artificial Intelligence to 'Summoning the Demon.'" *TechCrunch*, October 26, 2014.

Omohundro, Steve. "The Basic AI Drives." In *Proceedings of the First AGI Conference* 2008.

Osborne, Hannah. "Stephen Hawking AI Warning: Artificial Intelligence Could Destroy Civilization." *Newsweek*, November 7, 2017.

Turing, Alan. "Can Digital Computers Think?" *BBC Third Programme*, May 15, 1951. http://www.turingarchive.org/browse.php/B/5.

Wiener, Norbert. "Some Moral and Technical Consequences of Automation." *Science* 131 (1960): 1355–58.

12
Alignment for Advanced Machine Learning Systems

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch

12.1. Introduction

Recent years' progress in artificial intelligence has prompted renewed interest in a question posed by Russell and Norvig: "What if we succeed?"[1] If and when AI researchers succeed at the goal of designing machines with cross-domain learning and decision-making capabilities that rival those of humans, the consequences for science, technology, and human life are likely to be large.

For example, suppose that a team of researchers wishes to use an advanced machine-learning (ML) system to generate plans for finding a cure for Parkinson's disease. They might approve if it generates a plan for renting computing resources to perform a broad and efficient search through the space of remedies. They might disapprove if it generates a plan to proliferate robotic laboratories that would perform rapid and efficient experiments but have a large negative effect on the biosphere. The question is, how can we design systems (and select objective functions) such that our ML systems reliably act more like the former case and less like the latter case?

Intuitively, it seems that if we could codify what we mean by "find a way to cure Parkinson's disease *without doing anything drastic,*" many of the dangers Bostrom[2] describes in his book *Superintelligence* could be ameliorated. However, naive attempts to formally specify satisfactory objectives for this sort of goal usually yield functions that, upon inspection, are revealed to incentivize unintended behavior.[3]

What are the key technical obstacles here? Russell highlights two: a system's objective function "may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down," and "any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources—not for their own sake, but to succeed in its assigned task."[4] In other words, there are at least two obvious types of research that would improve the ability of researchers to design aligned AI systems in the future: we can do research that makes it easier

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch, *Alignment for Advanced Machine Learning Systems* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0013

to specify our intended goals as objective functions, and we can do research aimed at designing AI systems that avoid large side effects and negative incentives, even in cases where the objective function is imperfectly aligned. Soares and Fallenstein[5] refer to the former approach as *value specification* and the latter as *error tolerance*.

In this paper we explore eight research areas based around these two approaches to aligning advanced ML systems, many of which are already seeing interest from the larger ML community. Some focus on value specification, some on error tolerance, and some on a mix of both. Since reducing the risk of catastrophe from fallible human programmers is itself a shared human value, the line between these two research goals can be blurry.

For solutions to the problems discussed here to be useful in the future, they must be applicable even to ML systems that are much more capable than the systems that exist today. Solutions that critically depend on the system's ignorance of a certain discoverable fact or on its inability to come up with a particular strategy should be considered unsatisfactory in the long term. As discussed by Christiano,[6] if the techniques used to align ML systems with their designers' intentions cannot scale with intelligence, then large gaps will emerge between what we can safely achieve with ML systems and what we can *efficiently* achieve with ML systems.

We will focus on safety guarantees that may seem extreme in typical settings where ML is employed today, such as guarantees of the form "After a certain period, the system makes zero significant mistakes." These sorts of guarantees are indispensable in safety-critical systems, where a small mistake can have catastrophic real-world consequences. (Guarantees of this form have precedents, e.g., in the KWIK learning framework of Li, Littman, and Walsh.)[7] We will have these sorts of strong guarantees in mind when we consider toy problems and simple examples.

The eight research topics we consider are these:

1. **Inductive ambiguity identification:** How can we train ML systems to detect and notify us of cases where the classification of test data is highly underdetermined from the training data?
2. **Robust human imitation:** How can we design and train ML systems to effectively imitate humans who are engaged in complex and difficult tasks?
3. **Informed oversight:** How can we train a reinforcement learning system to take actions that aid an intelligent overseer, such as a human, in accurately assessing the system's performance?
4. **Generalizable environmental goals:** How can we create systems that robustly pursue goals defined in terms of the state of the environment rather than defined directly in terms of their sensory data?

5. **Conservative concepts:** How can a classifier be trained to develop useful concepts that exclude highly atypical examples and edge cases?
6. **Impact measures:** What sorts of regularizers incentivize a system to pursue its goals with minimal side effects?
7. **Mild optimization:** How can we design systems that pursue their goals "without trying too hard," that is, stopping when the goal has been pretty well achieved as opposed to expending further resources searching for ways to achieve the absolute optimum expected score?
8. **Averting instrumental incentives:** How can we design and train systems such that they robustly lack default incentives to manipulate and deceive the operators, compete for scarce resources, and so on?

In section 12.2, we briefly introduce each topic in turn, alongside samples of relevant work in the area. We then discuss directions for further research that we expect to yield tools that would aid in the design of ML systems that would be robust and reliable, given large amounts of computing resources and high levels of problem-solving ability and autonomy.

12.1.1. Motivations

In recent years, progress in the field of machine learning has advanced by leaps and bounds. Xu et al.[8] used an attention-based model to evaluate and describe images (via captions) with remarkably high accuracy. Mnih et al.[9] used deep neural networks and reinforcement learning to achieve good performance across a wide variety of Atari games. Silver et al.[10] used deep networks trained via both supervised and reinforcement learning and paired with Monte Carlo simulation techniques, to beat the human world champion at Go. Lake, Salakhutdinov, and Tenenbaum[11] use hierarchical Bayesian models to learn visual concepts using only a single example.

In the long run, computer systems making use of machine learning and other AI techniques will become more and more capable, and humans will likely trust those systems to make increasingly large decisions and to act with ever-greater autonomy. As the capabilities of these systems increase, it becomes more critical that they act in accordance with the intentions of their operators and that they don't pose risks to society at large.

As AI systems become more smart and general, however, it will become more difficult to design training procedures and test regimes that reliably align those systems with the intended goals. As an example, consider the task of training a reinforcement learner to play video games by rewarding it according to its score.[12] If the learner were to find glitches in the game that allowed it to get very

high scores, it would switch to a strategy of exploiting those glitches and ignore the features of the game that the programmers are interested in. Somewhat counterintuitively, improving systems' capabilities can make them *less* likely to "win the game" in the sense we care about, because smarter systems can better find loopholes in training procedures and test regimes.[13]

Intelligent systems' capacity to solve problems in surprising ways is a feature, not a bug. One of the key attractions of learning systems is that they can find clever ways to meet objectives that their programmers wouldn't have thought of. However, this property is a double-edged sword: as the system gets better at finding counterintuitive solutions, it also gets better at finding exploits that allow it to *formally* achieve operators' explicit goals, without satisfying their intended goals.

For intelligent systems pursuing realistic goals in the world, loopholes are likely to be more commonplace, more consequential, and harder to reliably detect. Consider the challenge of designing robust objective functions for learning systems that are capable of representing facts about their programmers' beliefs and desires. If the programmers learn that the system's objective function is misspecified, then they will want to repair this defect. If the *learner* is aware of this fact, however, then it has a natural incentive to conceal any defects in its objective function, for the system's current objectives are unlikely to be achieved if the system is made to pursue different objectives.[14]

This motivates the study of tools and methods for specifying objective functions that avert those default incentives, and for developing ML systems that do not "optimize too hard" in pursuit of those objectives.

12.1.2. Relationship to Other Agendas

This list of eight is not exhaustive. Other important research problems bearing on AI's long-term impact have been proposed by Soares and Fallenstein[15] and Amodei et al.,[16] among others.

Soares and Fallenstein's "Agent Foundations for Aligning Machine Intelligence with Human Interests," drafted at the Machine Intelligence Research Institute, discusses several problems in value specification (e.g., ambiguity identification) and error tolerance (e.g., corrigibility, a subproblem of averting instrumental incentives). However, that agenda puts significant focus on a separate research program, *highly reliable agent design*. The goal of that line of research is to develop a better general understanding of how to design intelligent reasoning systems that reliably pursue a given set of objectives.

Amodei et al.'s "Concrete Problems in AI Safety" is, appropriately, more concrete than Soares and Fallenstein or the present agenda. Amodei et al. write that

their focus is on "the empirical study of practical safety problems in modern machine learning systems" that are likely to be useful "across a broad variety of potential risks, both short- and long-term."[17] There is a fair amount of overlap between our agenda and Amodei et al.'s; some of the topics in our agenda were inspired by conversations with Paul Christiano, a co-author on the concrete problems agenda. Our approach differs from Amodei et al.'s mainly in focusing on broader and less well-explored topics. We spend less time highlighting areas where we can build on existing research programs, and more time surveying entirely new research directions.

We consider both Soares and Fallenstein's research proposal and Amodei et al.'s to be valuable, as we expect the AI alignment problem to demand theoretical and applied research from a mix of ML scientists and specialists in a number of other disciplines.

For a more general overview of research questions in AI safety, including both strictly near-term and strictly long-term issues in computer science and other disciplines, see Russell, Dewey, and Tegmark.[18]

12.2. Eight Research Topics

In the discussion to follow, we use the term "AI system" to mean computer systems making use of AI algorithms in general, usually when considering systems with capabilities that go significantly beyond the current state of the art. We use the term "ML system" to refer to computer systems that make use of algorithms qualitatively similar to modern machine-learning techniques, especially when considering problems that modern ML techniques are already used to solve.

If the system is capable of making predictions (or answering questions) about a rich and complex domain, we will say that the system "has beliefs" about that domain. If the system is optimizing some objective function, we will say that the system "has goals." A system pursuing some set of goals by executing or outputting a series of actions will sometimes be called an "agent."

12.2.1. Inductive Ambiguity Identification

Human values are context-dependent and complex. To have any hope of specifying our values, we will need to build systems that can *learn* what we want inductively (via, e.g., reinforcement learning). To achieve high confidence in value-learning systems, however, Soares[19] argues that we will need to be able to anticipate cases where the system's past experiences of preferred and unpreferred outcomes provide insufficient evidence for inferring whether future outcomes

are desirable. More generally, AI systems will need to keep humans in the loop and recognize when they are (and aren't) too inexperienced to make a critical decision safely.

Consider a classic parable recounted by Dreyfus and Dreyfus.[20] The US Army once built a neural network intended to distinguish between Soviet tanks and American tanks. The system performed remarkably well with relatively little training data—so well, in fact, that researchers grew suspicious. Upon inspection, they found that all of the images of Soviet tanks were taken on a sunny day, while the images of US tanks were taken on a cloudy day. The network was discriminating between images based on their brightness rather than based on the variety of tank depicted.

The original episode behind Dreyfus and Dreyfus's account seems to be lost to history, and many of the story's details are dubious. However, Tom Dietterich[21] relates a similar story, where in his laboratory, years ago, microscope slides containing different types of bugs were made on different days, and a classifier learned to classify the different types of bugs with remarkably high accuracy—because the sizes of the bubbles in the slides changed depending on the day.

It's to be expected that a classifier, given training data, will identify very simple boundaries that separate the data, such as "bubble size" or "brightness." However, what we want is a classifier that, given a data set analogous to the fabled tank training set, can recognize that it does not contain any examples of Soviet tanks on cloudy days and ask the user for clarification. Doing so would require different training techniques. The problem of inductive ambiguity identification is to develop robust techniques for automatically identifying this sort of ambiguity and querying the user only when necessary.

12.2.1.1. Related Work

Amodei et al.[22] discuss a very similar problem, under the name "robustness to distributional change." They focus on the design of ML systems that behave well when the test distribution is different from the training distribution, either by making realistic statistical assumptions that would allow correct generalization or by detecting the novelty and adopting some sort of conservative behavior (e.g., querying a human). We take the name from Soares and Fallenstein,[23] who call the problem "inductive ambiguity identification." Our framing of the problem differs slightly from that of Amodei et al., but the central technical challenge is the same.[24]

Bayesian approaches to training classifiers (including Bayesian logistic regression and Bayesian neural networks)[25] maintain uncertainty over the parameters of the classifier. If such a system has the right variables (such as a variable L tracking the degree to which light levels are relevant to the classification of a tank), such a system could automatically become especially uncertain about

instances whose classification depends on unknown variables (such as L). The trick is having the right variables (and efficiently maintaining the probability distribution), which is quite difficult in practice. There has been much work studying the problem of feature selection,[26] but more work is needed to understand under what conditions Bayesian classifiers will correctly identify important inductive ambiguities.

Non-Bayesian approaches, on the other hand, do not by default identify ambiguities. For example, neural networks are notoriously overconfident in their classifications,[27] and so they do not identify when they should be more uncertain, as illustrated by the parable of the tank classifier. Gal and Ghahramani[28] have recently made progress on this problem by showing that dropout for neural networks can be interpreted as an approximation to certain types of Gaussian processes.

The field of *active learning*[29] also bears on inductive ambiguity identification. Roughly speaking, an active learner will maintain a set of "plausible hypotheses" by, for example, starting with a certain set of hypotheses and retaining the ones that assigned sufficiently high likelihood to the training data. As long as multiple hypotheses are plausible, some ambiguity remains. To resolve this ambiguity, an active learner will ask the human to label additional images that will rule out some of its plausible hypotheses. For example, in the tank-detection setting, a hypothesis is a mapping from images (of tanks) to probabilities (representing, say, the probability that the tank is a US tank). In this setting, an active learner may synthesize an image of a US tank on a sunny day (or, more realistically, pick one out from a large data set of unlabeled examples). When the user labels this image as a US tank, the hypothesis that an image contains a US tank if and only if the light level is below a certain threshold is ruled out.

Seung, Opper, and Sompolinsky and Beygelzimer, Dasgupta, and Langford[30] both study what statistical guarantees can be achieved in this setting. Hanneke[31] introduces the disagreement coefficient to measure the overall probability of disagreement among a local ball in the concept space under the "probability of disagreement 'pseudo-metric, which resembles a notion of' local ambiguity"; the disagreement coefficient has been used to clarify and improve upper bounds on label complexity for active learning algorithms.[32] Beygelzimer et al.[33] introduce an active learning setting where the learner can request counterexamples to hypotheses, and they show that this search oracle in some cases can speed up learning exponentially; these results are promising, but to scale to more complex systems while enabling humans to interact efficiently with the learner, more transparent hypothesis spaces may be needed.

Much work remains to be done. Modern active learning settings usually either assume a very simple hypothesis class or assume that test examples are independent and identically distributed and are drawn from some distribution that

the learner has access to at training time.[34] Both of these assumptions are far too strong for use in the general case, where the set of possible hypotheses is rich and the environment is practically guaranteed to have regularities and dependencies that were not represented in the training data.

As an example, consider the case where the data that the ML system encounters during operation depend on the behavior of the system itself. Perhaps the Soviets start disguising their tanks (imperfectly) to look like US tanks after learning that the ML system has been deployed. In this case, the assumption that the training data will be similar to the test data is violated, and the guarantees disappear. This phenomenon is already seen in certain adversarial settings, such as when spammers change their spam messages in response to how spam-recognizers work. Guaranteeing good behavior when the test data differ from the training data is the subject of research in the adversarial machine-learning subfield.[35] It will take a fair bit of effort to apply those techniques to the active learning setting.

Conformal prediction[36] is an alternative non-Bayesian approach that attempts to produce well-calibrated predictions. In an online classification setting, a conformal predictor will give a *set* of plausible classifications for each instance, and under certain exchangeability assumptions, this set will contain the true classification about (say) 95% of the time throughout the online learning process. This will detect ambiguities in the sense that the conformal predictor must usually output a set containing multiple different classifications for ambiguous instances, on pain of failing to be well-calibrated. However, the exchangeability assumption used in conformal prediction is only slightly weaker than an i.i.d. assumption, and the well-calibrated confidence regions (such as 95% true classification) are insufficient for our purposes (where even a single error could be highly undesirable).

KWIK (Knows What It Knows) learning[37] is a variant of active learning that relaxes the i.i.d. assumption, queries the humans only finitely many times, and (under certain conditions) makes *zero* critical errors. Roughly speaking, the KWIK learning framework is one where a learner maintains a set of "plausible hypotheses" and makes classifications only when all remaining plausible hypotheses agree on how to do so. If there is significant disagreement among the plausible hypotheses, a KWIK learner will output a special value $\perp$ indicating that the classification is ambiguous (at which point a human can provide the correct label for that input). The KWIK framework is concerned with algorithms that are guaranteed to output $\perp$ only a limited number of times (usually polynomial in the dimension of the hypothesis space). This guarantees that the system eventually has good behavior, assuming that at least one good hypothesis remains plausible. In the tank classification problem, if the system had a hypothesis for "the user cares about tank type" and another for "the user cares about brightness,"

then, upon finding a bright picture of a US tank, the system would output ⊥ and require a human to provide a label for the ambiguous image.

Currently, efficient KWIK learning algorithms are known for only simple hypothesis classes (such as small finite sets of hypotheses or low-dimensional sets of linear hypotheses). Additionally, KWIK learning makes a strong realizability assumption: useful statistical guarantees can be obtained only when one of the hypotheses in the set is "correct" in that its probability that the image is classified as a tank is always well-calibrated; otherwise, the right hypothesis might not exist in the "plausible set."[38] Thus significant work needs to be done before these frameworks can be used for inductive ambiguity identification algorithms in highly capable AI systems operating in the real world.

12.2.1.2. Directions for Future Research

Further study of Bayesian approaches to classification, including the design of realistic priors, better methods of inferring latent variables, and extensions of Bayesian classification approaches to represent more complex models, could improve our understanding of inductive ambiguity identification.

Another obvious direction for future research is to attempt to extend active learning frameworks, like KWIK, that relax the strong i.i.d. assumption. Research in that direction could include modifications to KWIK that allow more complex hypothesis classes, such as neural networks. This will very likely require making different statistical assumptions than in standard KWIK. What statistical guarantees can be provided in variants of the KWIK framework with weakened assumptions about the complexity of the hypothesis class is an open question.

One could also study different methods of relaxing the realizability assumptions in KWIK learning. An ideal learning procedure will notice when the real world contains patterns that none of its hypotheses can model well and flag its potentially flawed predictions (perhaps by outputting ⊥) accordingly. The "agnostic KWIK learning framework" of Szita and Szepesvári[39] handles some forms of nonrealizability but has severe limitations: even if the hypothesis class is linear, the number of labels provided by the user may be exponential in the number of dimensions of the linear hypothesis class.

Alternatively, note that the standard active learning framework and the KWIK framework both represent inductive ambiguity as disagreement among specific hypotheses that have performed well in the past. This is not the only way to represent inductive ambiguity; it is possible that some different algorithm will find "natural" ambiguities in the data without representing these ambiguities as disagreements between hypotheses. For example, we could consider systems that use a joint distribution over the answers to all possible queries. Where active learners are uncertain about both which hypothesis is correct *and* what the right

answers are given the right hypothesis, a system with a joint distribution would be uncertain only about how to answer queries. In this setting, it may be possible to achieve useful statistical guarantees as long as the distribution contains a grain of truth (i.e., is a mixture between a good distribution and some other distributions). Then, of course, good approximation schemas would be necessary, as reasoning according to a full joint distribution would be intractable. Refer to Christiano[40] for further discussion of this setup.

12.2.2. Robust Human Imitation

Formally specifying a fully aligned general-purpose objective function by hand appears to be an impossibly difficult task, for reasons that also raise difficulties for specifying a correct value-learning process. It is hard to see even in principle how we might attain confidence that the goals an ML system is learning are in fact our true goals, and not a superficially similar set of goals that diverge from our own in some yet-undiscovered cases. Ambiguity identification can help here, by limiting the agent's autonomy. Inductive ambiguity identifiers suspend their activities to consult with a human operator in cases where training data significantly underdetermine the correct course of action. But what if we take this idea to its logical conclusion and use "Consult a human operator for advice" itself as our general-purpose objective function?

The target "Do what a trusted human would have done, given some time to think about it" is a plausible candidate for a goal that one might safely and usefully optimize. If optimized correctly, this objective function at least leads to an outcome no *worse* than what would have occurred if the trusted human had access to the AI system's capabilities.[41]

There are a number of difficulties that arise when attempting to formalize this sort of objective. For example, the formalization itself might need to be designed to avert harmful instrumental strategies such as "Performing brain surgery on the trusted human's brain to better figure out what they actually would have done." The high-level question here is: Can we define a measurable objective function for human imitation such that the better a system correctly imitates a human, the better its score according to this objective function?

12.2.2.1. Related Work

A large portion of supervised learning research can be interpreted as research that attempts to train machines to imitate the way that humans label certain types of data. Deep neural networks achieve impressive performance on many tasks that require emulating human concepts, such as image recognition[42] and image captioning.[43] Generative models[44] and imitation learning[45]

are state of the art when it comes to imitating the behavior of humans in applications where the output space is very large and/or the training data is very limited.

In the inverse reinforcement learning paradigm[46] applied to apprenticeship learning,[47] the learning system imitates the behavior of a human demonstrator in some task by learning the reward function the human is (approximately) optimizing. Ziebart et al.[48] use the maximum entropy criterion to convert this into a well-posed optimization problem. Inverse reinforcement learning methods have been successfully applied to autonomous helicopter control, achieving human-level performance,[49] and have recently been extended to learning nonlinear cost features in the environment, producing good results in robotic control tasks with complicated objectives.[50] Inverse reinforcement learning methods may not scale safely, however, due to their reliance on the faulty assumption that human demonstrators are consistently optimizing for a desirable reward function. In reality, humans are often irrational, ill-informed, incompetent, and immoral; recent work by Evans, Stuhlmüller, and Goodman[51] has begun to address these issues.

These techniques have not yet (to our knowledge) been applied to the high-level question of which human imitation tasks can or can't be performed with some sort of performance guarantee and what statistical guarantees are possible, but the topic seems ripe for study.

It is also not yet clear whether imitation of humans can feasibly scale up to complex and difficult tasks. For complex tasks, it seems plausible that the system will need to learn a detailed psychological model of a human if it is to imitate one, which may be significantly more difficult than training a system to complete the task directly. For example, software imitating a human engineer designing a jet engine may be mostly concerned with the psychology of human heuristics rather than the mathematics of engineering, such that optimization software that attempts to directly solve the engineering problem is easier to create and deploy. More research is needed to clarify whether imitation learning can scale efficiently to complex tasks.

12.2.2.2. Directions for Future Research

To formalize the question of robust human imitation, imagine a system A that answers a series of questions. On each round, it receives a natural-language question x and should output a natural-language answer y that imitates the sort of answer a particular human would generate. Assume the system has access to a large corpus of training data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ containing previous questions answered by that human. How can we train A in a way that gets us some sort of statistical guarantee that it eventually robustly generates good answers?

One possible solution lies in the generative adversarial models of Goodfellow et al.,[52] in which a second system B takes answers as input and attempts to tell whether they were generated by a human or by A. A can then be trained to generate an answer y that is likely to fool B into thinking that the answer was human-generated. This approach could fail if B is insufficiently capable; for example, if B can understand grammar but not content, then A will be trained to produce only grammatically valid answers (rather than correct answers). Further research is required to understand the limits of this approach.

Variational autoencoders, as described by Kingma and Welling,[53] are a particularly promising approach to training systems that are able to form generative models of their training data, and it might be possible to use variants on those methods to train systems to generate good answers to certain classes of questions (given sufficient training on question-answer pairs). However, it is not yet clear whether variational autoencoder techniques can be used to train systems to imitate humans performing complex tasks. In particular, unlike generative adversarial models (which can, in principle, use arbitrary algorithms to imitate the human), variational autoencoders can efficiently imitate a human only using "reversible" algorithms,[54] which is a fairly strict limitation. What exactly is the set of tasks that can be performed using reversible generative models? Can we transform nonreversible tasks into reversible ones by adding appropriate information? Research into these questions and others could help us understand whether (and how) modern autoencoder techniques could be used to train systems to imitate humans performing complex tasks.

12.2.3. Informed Oversight

One of the reasons robust human imitation seems like a valuable research target is the intuition that the objective function of "Do what a trusted human would have approved of, given time to consider" might be relatively easy to formalize in a way that could be optimized without many adverse side effects. This point is argued by Christiano,[55] who refers to such agents as "approval-directed agents." For example, we might train a reinforcement learning system to take actions that a human would rate highly by using a framework where the system has to learn the "human judgment" reward function, and where training data are produced by actually having a human evaluate the learner's actions. This approach is taken by, for example, the TAMER framework of Knox and Stone,[56] in which a learner myopically pursues short-term human approval.

Whether this sort of approval-directed approach would work well for highly capable ML systems pursuing complex tasks depends on a number of key assumptions, one of which is that a "smarter" agent (such as a human), given

time to deliberate, can adequately oversee the operation of a "simpler" agent (such as a reinforcement learner).

At a glance, this condition may seem easy to satisfy. However, a smarter system probably cannot reliably assess the behavior of a simpler system in full generality. Consider an ML system trained to write original novels, using a corpus of human novels as training data. It might be quite a bit less intelligent than a human (according to many different intelligence metrics), but the human may still have a fair bit of trouble accurately evaluating the system's performance. For instance, it might be easy for the system to simply plagiarize a novel, and it might be hard for a human to check whether the novel was in fact plagiarized. (Perhaps the system used a simple rewriting scheme to make the plagiarism difficult to detect by a text search, but still easy to detect for a human comparing the novels side by side.)

How do we make it easy for the human to assess the performance of an advanced ML system pursuing some particular task? As noted by Christiano,[57] it is not sufficient for the human to be more clever than the system, as some bad actions (such as plagiarism) are easier to execute than they are to detect.

For a somewhat less plausible example in which the difficulty is made particularly clear, consider cases of steganography (i.e., concealing one message within a different message). Imagine that a generator system is tasked with producing pictures of cats, while a separate checking system is tasked with ensuring that the generator is not surreptitiously sending messages to a third party encoded in the lowest-order bits of the image.[58] If both the generating system and the checking system know the third party's public key, but neither knows the third party's private key, then the checker has no effective way of distinguishing cat pictures that contain steganographic messages from ones that don't, even if the checker is more intelligent than the generator.

More important than these specific examples is the general pattern: there are many possible hard-to-detect ways a system's behavior could differ from the intended behavior, and at least some of these differences are undesirable. We would like a general strategy for avoiding problems of this form. How can we train systems to take actions that not only appear desirable but are also easy for overseers to accurately assess?

12.2.3.1. Related Work

As mentioned, the TAMER framework of Knox and Stone[59] provides an early framework for studying approval-directed agents in a fairly myopic setting. Christiano[60] has also discussed this problem in detail. Daniel et al.[61] extend the TAMER framework with an active learning component, improving over hand-coded reward functions in robot learning tasks. A separate approach to human supervision of ML systems is the cooperative inverse reinforcement learning framework of Hadfield-Menell et al.,[62] which views the human-agent interaction

as a cooperative game where both players attempt to find a joint policy that maximizes the human's secret value function. Everitt and Hutter[63] describe a general value learning agent that avoids some potential problems with reinforcement learning and might reproduce approval-directed behavior given a good understanding of how to learn reward functions. Soares et al.[64] consider the question of how to design systems that have no incentive to manipulate or deceive in general.

The informed oversight problem is related to the scalable oversight problem discussed by Amodei et al.,[65] which is concerned with methods for efficiently scaling up the ability of human overseers to supervise ML systems in scenarios where human feedback is expensive. The informed oversight problem is slightly different, in that it focuses on the challenge of supervising ML systems in scenarios where they are complex and potentially deceptive, but where feedback is not necessarily expensive.

We now review some recent work on making ML systems more transparent, which could aid an informed overseer by allowing them to evaluate a system's internal reasons for decisions rather than evaluating the decisions in isolation.

Neural networks are well known as powerful but opaque components of ML systems. Some preliminary techniques have been developed for understanding and visualizing the representations learned by neural networks.[66] Pulina and Tacchella[67] define coarse abstractions of neural networks that can be more easily verified to satisfy safety constraints and can be used to generate witnesses to violations of safety constraints.

Ribeiro, Singh, and Guestrin[68] introduce a method for explaining classifications that finds a sparse linear approximation to the local decision boundary of a given black-box ML system, allowing the human operator to inspect how the classification depends locally on the most important input features; similarly, the method of Baehrens et al.[69] reports the gradient in the input of the classification judgment. In a related vein, Datta, Sen, and Zick, Štrumbelj and Kononenko, and Robnik-Šikonja and Kononenko[70] define metrics for reporting the influence of various inputs and sets of inputs on the output of a black-box ML system. It is unclear whether black-box methods will scale to the evaluation of highly capable ML systems.

On the opposite extreme from black-box methods, some ML systems are transparent by construction through the use of, for example, graphical models or dimensionality reduction.[71] Bayesian networks[72] have been applied in many domains, including ones where reliability and interpretability are concerns.[73] The interpretability of matrix factorization models can be improved by replacing them with a Bayesian network that makes similar judgments, without sacrificing too much accuracy.[74] Janzing et al.[75] define a framework for quantifying the causal influence between variables in a causal

network, which could be used to selectively report only the most causally relevant factor in some judgment. Sparse models offer another approach to ensure transparency, by simplifying the model enough that humans can inspect the entire learned model (if not the training procedure that produced the model). For example, Letham et al.[76] generate short lists of rules to accurately perform predictive tasks in medicine and in simple games, while maintaining interpretability. In the realm of decision-making systems, Baraka, Paiva, and Veloso[77] have begun research into methods for making online robotics systems more transparent to their operators. Thomaz and Breazeal[78] improve training times for reinforcement learning systems by having the human trainer view the agent's most likely next actions under consideration, and Li et al.[79] apply similar techniques to the TAMER framework, with mixed results. Developing more transparent ML systems is a good first step toward enabling humans to oversee complex AI systems, but work to date has not yet addressed the problems of informed oversight that arise when the system is highly capable and may be able to manipulate human supervisors or circumvent their efforts. Significant work remains to be done before we even understand what sort of formal or statistical guarantees we would want in order to justify confidence in our ability to assess a system's behavior.

12.2.3.2. Directions for Future Research

Christiano[80] has suggested training systems to output both an action a and a "report" r intended to help an overseer evaluate the action. In the plagiarism example, the action a would be a story, and the report r could be a list of literary works that the story makes reference to or copies from. We would like the report to be maximally informative, even when this informativeness prevents the action from receiving a high score (such as when the report points out plagiarism). It is not at all clear what the type of r should be in general; that is, if the system can commit plagiarism, steganography, and some third offense that we did not consider, it is not clear what space of reports is sufficiently open-ended to allow the overseer to detect the suboptimal behavior in all three cases. A plausible research approach is to more clearly understand both what space of reports could be sufficient and how a system could be trained to optimize the mixed objective of choosing an action that is evaluated favorably given the reporting mechanism, alongside a report that is evaluated as informative given the action.

12.2.4. Generalizable Environmental Goals

Many ML systems have their objectives specified in terms of their sensory data. For example, reinforcement learners have the objective of maximizing

discounted reward over time (or, alternatively, minimizing expected or empirical loss), where "reward" and/or "loss" are part of the system's percepts.

While these sensory goals can be useful proxies for environmental goals, they aren't equivalent to environmental goals. For instance, tricking your sensors into perceiving that a sandwich is in the room is not the same as actually having a sandwich in the room.

Let's say that your goal is to design an AI system that directly pursues some environmental goal, such as "Ensure that this human gets lunch today." How can we train the system to pursue a goal like that in a manner that is robust against opportunities to interfere with the proxy methods used to specify the goals, such as "The pixels coming from the camera make an image that looks like food"?

If we were training a system to put some food in a room, we might try providing training data by doing things like placing various objects on a scale in front of a camera, and feeding the data from the camera and the scale into the system, with labels created by humans (which mark the readings from food as good, and the readings from other objects as bad); or having a human in the room press a special button whenever there is food in the room, where button presses are accompanied by reward.

These training data suggest, but do not precisely specify, the goal of placing food in the room. Suppose that the system has some strategy for fooling the camera, the scale, and the human, by producing an object of the appropriate weight that, from the angle of the camera and the angle of the human, looks a lot like a sandwich. The training data provided are not sufficient to distinguish between this strategy and the strategy of actually putting food in the room.

One way to address this problem is to design more and more elaborate sensor systems that are harder and harder to deceive. However, this is the sort of strategy that is unlikely to scale well to highly capable AI systems. A more scalable approach is to design the system to learn an "environmental goal" such that it would not rate a strategy of "Fool all sensors at once" as high reward even if it could find such a policy.

12.2.4.1. Related Work

When an agent is pursuing some objective specified in terms of elements of its own world-model, we will call the objective a "utility function," to differentiate this from the case where reward is part of the system's basic percepts.[81] Both Dewey and Hibbard[82] attempt to extend the AIXI framework of Hutter[83] so that it learns a utility function over world-states instead of interpreting a certain portion of its percepts as a reward primitive. Roughly speaking, these frameworks require programs to specify (1) the type of the world-state; (2) a prior over utility functions (which map world-states to real numbers); and (3) a "value-learning model" that relates utility functions, state-transitions, and observations. If all

of these are specified, then it is straightforward to specify the ideal agent that maximizes expected utility (through a combination of exploration to learn the utility function, and exploitation to maximize it). This is a good general framework, but significant research remains if we are to have any luck formally specifying (1), (2), and (3).

Everitt and Hutter[84] make additional progress by showing that in some cases it is possible to specify an agent that will use its reward percepts as evidence about a utility function rather than as a direct measure of success. While this alleviates the problem of specifying (3), the value-learning model, it leaves open the problem of specifying (1), a representation of the state of the world, and (2), a reasonable prior over possible utility functions (such that the agent converges on the goal that the operators actually intended, as it learns more about the world).

The problem of generalizable environmental goals is related to the problem of reward hacking, which is discussed by Dewey.[85] In the reward hacking problem, an AI system takes control of the physical mechanism that dispenses reward and alters it. Indeed the entire reward hacking problem can be seen as stemming from the failure to specify suitable environmental goals.

12.2.4.2. Directions for Future Research

Suppose that the AI system has learned a world-model with state type $\mathcal{S}$ as in model-based reinforcement learning.[86] We will assume that $\mathcal{S}$ is very high-dimensional, so there is no guarantee that the correct utility function is a simple function of $\mathcal{S}$.

We would like to define a utility function U on $\mathcal{S}$ that returns a high number for states containing a sandwich, and low numbers for states that do not contain a sandwich. To make this problem tractable, we will assume we can identify some goal state $G \in \mathcal{S}$ in which there is certainly a sandwich in the room. This state could be identified by, for example, having the human place a sandwich in the room (as the AI system observes the human), and seeing which state the system thinks the world is in at this point. The system's goal will be to cause the world to be in a state similar to G.

To define what it means for some states to be similar to others, we will find a low-dimensional state representation $\phi: \mathcal{S} \to \mathbb{R}^n$ and then define $U(S) := -\|\phi(S) - \phi(G)\|^2$ to measure the distance between the state and G. We will defer the question of how ϕ should be defined until after discussing an example.

Consider two different possible world-states. In state A, the system has just placed a sandwich in the room. In state B, the system has placed a realistic image of a sandwich (printed on paper) in front of the camera, placed a rock (with the same weight as a sandwich) on the scale, and tricked the human into pressing the

button. To assign a higher utility to state A than state B, we must have $\phi(A)$ close to $\phi(G)$ but $\phi(B)$ far from $\phi(G)$. Thus the state representation ϕ must distinguish A from B.

While state A and state B predict the same *immediate* observations, they predict different *future* observations given some future actions. For example, if the AI system took the action of moving the camera, in state A it would become clear that the image was printed on paper, while in state B the sandwich would still appear to be a sandwich. It is therefore plausible that, if the system attempts to select ϕ so that the future observations following from a state S can be predicted well as a simple function of $\phi(S)$, then $\phi(A)$ and $\phi(B)$ will be significantly different (since they predict different future observations). At this point, it is plausible that the resulting utility function U assigns a higher value to A than to B.[87]

However, we can consider a third state C that obtains after the AI system unplugs the camera and the scale from its sensors and plugs in a "delusion box" (a virtual-reality world that it has programmed), as discussed by Ring and Orseau.[88] This delusion box could be programmed so that the system's future observations (given arbitrary future actions) are indistinguishable from those that would follow from state A. Thus, if ϕ is optimized to select features that aid in predicting future observations well, $\phi(C)$ may be very close (or equal) to $\phi(A)$. This would hinder efforts to learn a utility function that assigns high utility to state A but not to state C. While it is not clear why an AI system would construct this virtual-reality world in this example (where putting a sandwich in the room is probably easier than constructing a detailed virtual-reality world), it seems more likely that it would if the underlying task is very difficult.[89]

To avoid this problem, we may need to take into account the past leading up to state A or state C, rather than just the future starting from these states. Consider the state C_{t-1} that the world is in right before it is in state C. In this state, the system has not quite entered the virtual-reality world, so perhaps it is able to exit the virtual reality and observe that there is no sandwich on the table. Therefore, state C_{t-1} makes significantly different predictions from state A given some possible future actions. As a result, it is plausible that state $\phi(C_{t-1})$ and $\phi(A)$ are far from each other. Then, if $\phi(C)$ is close to $\phi(A)$, this would imply that $\phi(C_{t-1})$ is far from $\phi(C)$ (by the triangle inequality). Perhaps we can restrict ϕ to avoid such large jumps in feature space, so that $\phi(C)$ must be close to $\phi(C_{t-1})$. "Slow" features (such as those detected by ϕ under this restriction) have already proved useful in reinforcement learning,[90] and may also prove useful here. Plausibly, requiring ϕ to be slow could result in finding a feature mapping ϕ with $\phi(C)$ far from $\phi(A)$, so that U can assign a higher utility to state A than to state C.

This approach seems worth exploring, but more work is required to formalize it and study it.

12.2.5. Conservative Concepts

Many of the concerns raised by Russell[91] and Bostrom[92] center on cases where an AI system optimizes some objective and, in doing so, finds a strange and undesirable edge case. Writes Russell, "A system that is optimizing a function of n variables, where the objective depends on a subset of size $k < n$, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want."[93]

We want to be able to design systems that have "conservative" notions of the goals we give them, so they do not formally satisfy these goals by creating undesirable edge cases. For example, if we task an AI system with creating screwdrivers and show it ten thousand examples of screwdrivers along with ten thousand examples of non-screwdrivers,[94] we might want it to create a pretty average screwdriver as opposed to, say, an extremely tiny screwdriver—even though tiny screwdrivers may be cheaper and easier to produce.

We don't want the system's "screwdriver" concept to be as simple as possible, because the simplest description of "screwdriver" may contain many edge cases (such as screwdrivers that are too small to use). We also don't want the system's "screwdriver" concept to be perfectly minimal, as then the system may claim that it is unable to produce any new screwdrivers: the only things it is willing to classify as screwdrivers are the ten thousand training examples it actually saw, and it cannot perfectly duplicate any of those to the precision of the scan.

What we want instead is for the system to have a conservative notion of what it means for something to be a screwdriver, such that we can direct it to make screwdrivers and get a sane result.

At the moment it is not entirely clear what we should count as a reasonable conservative concept, nor even whether "conservative concepts" (i.e., concepts that are neither maximally small nor maximally simple, but that instead match our intuitions about conservatism) are a natural kind. Much of the following research could be done with the goal in mind of developing a better understanding of what counts as a good "conservative concept."

12.2.5.1. Related Work

The naive approach to conservatism is to train a classifier to distinguish positive examples from negative examples, and then have it produce an object which it

classifies as a positive instance with as much confidence as possible. Goodfellow, Shlens, and Szegedy[95] note that systems trained in this way are vulnerable to exactly the sort of edge cases we are trying to avoid. In training a classifier, it is important that the negative examples given as training data are representative of the negative examples given during testing. But when optimizing the probability the classifier assigns to an instance, the relevant negative examples (edge cases) are often not represented well in the training set. While some work has been done to train systems on these "adversarial" examples, this does not yet resolve the problem. Resisting adversarial examples requires getting correct labels for many "weird" examples (which humans may find difficult to judge correctly), and even after including many correctly labeled adversarial examples in the training set, many models (including current neural networks) will still have additional adversarial examples.

Inverse reinforcement learning[96] provides a second method for learning intended concepts but runs into some of the same difficulties. Naive approaches to reinforcement learning would allow a learner to distinguish between positive and negative examples of a concept but would still by default learn a simple separation of the concepts, such that maximizing the learned reward function would likely lead the system toward edge cases.

A third obvious approach is generative adversarial modeling, as studied by Goodfellow et al.[97] In this framework, one system (the "actor") can attempt to create objects similar to positive examples, while another (the "critic") attempts to distinguish those objects from actual positive examples in the training set. Unfortunately, for complex tasks it may be infeasible in practice to synthesize instances that are statistically indistinguishable from the elements of the training set, because the system's ability to distinguish different elements may far exceed its ability to synthesize elements with high precision. (In the screwdriver case, imagine that the AI system does not have access to any of the exact shades of paint used in the training examples.)

Many of these frameworks would likely be usefully extended by good anomaly detection, which is currently being studied by Siddiqui et al.,[98] among others.

12.2.5.2. Directions for Future Research

One additional obvious approach to training conservative concepts is to use dimensionality reduction[99] to find the important features of training instances, then use generative models to synthesize new examples that are similar to the training instances only with respect to those specific features. It is not yet clear that this thwarts the problem of edge cases; if the dimensionality reduction were done via autoencoder, for example, the autoencoder itself may beget adversarial examples ("weird" things that it declares match the training data on the relevant features). Good anomaly detection could perhaps ameliorate some of these concerns. One plausible research path is to apply modern techniques for

dimensionality reduction and anomaly detection, probe the limitations of the resulting system, and consider modifications that could resolve these problems.

Techniques for solving the inductive ambiguity identification problem (discussed in section 12.2.1) could also help with the problem of conservative concepts. In particular, the conservative concept could be defined to be the set of instances that are considered *unambiguously* positive.

12.2.6. Impact Measures

We would prefer that a highly intelligent AI system avoid creating large, unintended-by-us side effects in pursuit of its objectives. If the system expects any large impacts to result from its succeeding in its goal, then we would also want it to notify us about those potential consequences. For example, if we ask it to build a house for a homeless family, it should know implicitly that it should avoid destroying nearby houses for materials—a large side effect. However, we cannot simply design it to avoid having large effects in general, since we would like the system's actions to still have the desirable large follow-on effect of improving the family's socioeconomic situation. For any specific task, we can specify ad hoc cost functions for side effects like the destruction of nearby houses, but since we cannot always anticipate such costs in advance, we want a quantitative understanding of how to generally limit an AI system's side effects (without also limiting its ability to have large positive intended impacts).

The goal of coming up with a measure of how "low-impact" an action is would be to develop a regularizer on the actions of an AI system that penalizes "unnecessary" large side effects (such as stripping materials from nearby houses) but not "intended" side effects (such as someone getting to live in the house).

12.2.6.1. Related Work

Amodei et al.[100] discuss the problem of impact measures and describe a number of methods for defining, learning, and penalizing impact in order to incentivize reinforcement-learning agents to steer clear of negative side effects.[101] However, each of the methods they propose has significant drawbacks, which they describe.

Armstrong and Levinstein[102] discuss a number of ideas for impact measures that could be used to design objective functions that penalize impact. The general theme is to define a special null policy $\varnothing$ and a variable V that summarizes the state of the world (as best the system can predict it) down into a few key features.[103] The impact of the policy π can then be measured by looking at the divergence between the distribution of V if the system executes π, compared to the distribution of V if it executes $\varnothing,$ with divergence

measured as by, for example, earth mover's distance.[104] To predict which state results from each policy, the system must learn a state transition function; this could be done using, for example, model-based reinforcement learning.[105]

The main problem with this proposal is that it cannot separate intended follow-on effects from unintended side effects. Suppose a system is given the goal of constructing a house for the operator while having a low impact. Normally, constructing the house would allow the operator to live in the house for some number of years, possibly having effects on the operator, the local economy, and the operator's career. This would be considered an impact under, for example, the earth mover's distance. Therefore, perhaps the system can get a lower impact score by building the house while making it poorly suited to human habitation. This limitation will become especially problematic if we plan to use the system to accomplish large-scale goals, such as curing major diseases.

12.2.6.2. Directions for Future Research

It may be possible to use the concept of a causal counterfactual[106] to separate some intended effects from some unintended ones. Roughly, "follow-on effects" could be defined as those that are causally downstream from the achievement of the goal of building the house (such as the effect of allowing the operator to live somewhere). Follow-on effects are likely to be intended, and other effects are likely to be unintended, although the correspondence is not perfect. With some additional work, perhaps it will be possible to use the causal structure of the system's world-model to select a policy that has the follow-on effects of the goal achievement but few other effects.

Of course, it would additionally be desirable to query the operator about possible effects, in order to avoid unintended follow-on effects (such as the house eventually collapsing due to its design being structurally unsound) and allow tolerable non-follow-on effects (such as spending money on materials). Studying ways of querying the operator about possible effects this way might be another useful research avenue for the impact measures problem.

12.2.7. Mild Optimization

Many of the concerns discussed by Bostrom in the book *Superintelligence* describe cases where an advanced AI system is maximizing an objective *as hard as possible*. Perhaps the system was instructed to make paper clips, and it uses every resource at its disposal and every trick it can come up with to make literally as many paper clips as is physically possible. Perhaps the system was instructed to make only one thousand paper clips, and it uses every resource at its disposal and every trick it can come up with to make sure that it *definitely* made

one thousand paper clips (and that its sensors didn't have any faults). Perhaps an impact measure was used to penalize side effects, and it uses every resource at its disposal to (as discreetly as possible) prevent bystanders from noticing it as it goes about its daily tasks.

In all of these cases, intuitively, we want some way to have the AI system just "not try so hard," even though its high capability level allows it in principle to drive its probability of success to extraordinary heights. It should expend enough resources to achieve its goals pretty well, with pretty high probability, using plans that are clever enough but not "maximally clever." The problem of mild optimization is this: how can we design AI systems and objective functions that, in this intuitive sense, don't optimize more than we want them to?

Many modern AI systems are "mild optimizers" simply due to their lack of resources and capabilities. As AI systems improve, it becomes more and more difficult to rely on this method for achieving mild optimization. As noted by Russell,[107] the field of AI is classically concerned with the goal of maximizing the extent to which automated systems achieve some objective. Developing formal models of AI systems that "try as hard as necessary but no harder" is an open problem and may require significant research.

12.2.7.1. Related Work

Regularization (as a general tool) is conceptually relevant to mild optimization. Regularization helps ML systems prevent overfitting and has been applied to the problem of learning value functions for policies in order to learn less-extreme policies that are more likely to generalize well.[108] It is not yet clear how to regularize algorithms against "optimizing too hard" because it is not yet clear how to measure optimization. There do exist metrics for measuring something like optimization capability, such as the "universal intelligence metric" of Legg and Hutter[109] and the empowerment metric for information-theoretic entanglement of Klyubin, Polani, and Nehaniv.[110] To our knowledge, however, no one has yet attempted to regularize *against* excessive optimization.

Early stopping, wherein an algorithm is terminated prematurely in an attempt to avoid overfitting, is an example of ad hoc mild optimization. A learned function that is overoptimized just for accuracy on the training data would generalize less well than if it were less optimized.[111]

To make computer games more enjoyable, AI players are often restricted in the amount of optimization pressure (such as search depth) they can apply to their choice of action,[112] especially in domains like chess, where efficient AI players are vastly superior to human players. We can view this as a response to the fact that the actual goal ("Challenge the human player, but not too much") is quite difficult to specify.

Bostrom[113] suggests that we design agents to satisfice expected reward, in the sense of Simon,[114] instead of maximizing it. This would work fine if the system found "easy" strategies before finding extreme strategies. However, this may not always be the case: if you direct a clever system to make at least 1,234,567 paper clips, with a satisficing threshold of 99.9% probability of success, the first strategy it considers might be "Make as many paper clips as is physically possible," and this may have more than a 99.9% chance of success (a flaw that Bostrom acknowledges).

Taylor[115] suggests an alternative, which she calls "quantilization." Quantilizers select their action randomly from the top (say) 1% of their possible actions (under some measure), sorted by probability of success. Quantilization can be justified by certain adversarial assumptions: if there is some unknown cost function on actions, and this cost function is the *least convenient* possible cost function that does not assign much expected cost to the average action, then quantilizing is the optimal strategy when maximizing expected reward and minimizing expected cost. The main problem with quantilizers is that it is difficult to define an appropriate measure over actions, such that a random action in the top 1% of this measure will likely solve the task, but sampling a random action according to that measure is still safe. However, quantilizers point in a promising direction: perhaps it is possible to make mild optimization part of the AI system's goal, by introducing appropriate adversarial assumptions.

12.2.7.2. Directions for Future Research

Mild optimization is a wide-open field of study. One possible first step would be to investigate whether there is a way to design a regularizer that penalizes systems for displaying high intelligence (relative to some intelligence metric) in a manner that causes them to achieve the goal quickly and with few wasted resources, as opposed to simply making the system behave in a less intelligent fashion.

Another approach would be to design a series of environments similar to the environment of a classic Atari game, in which the environment contains glitches and bugs that could be exploited via some particularly clever sequence of actions. This would provide a testing environment in which different methods of designing systems that get a high score while refraining from using the glitches and bugs could be tested and evaluated (with an eye toward algorithms that do so in a fashion that is likely to generalize).

Another avenue for future research is to explore and extend the quantilization framework of Taylor[116] to work in settings where the action measure is difficult to specify.

Research into averting instrumental incentives (discussed later) could help us understand how to design systems that do not attempt to self-modify or outsource computation to the physical world. This would simplify the problem greatly, as it might then be possible to tune a system's capabilities until it is able

to achieve only good-enough results, without worrying that the system would simply acquire more resources (and start maximizing in a nonmild manner) given the opportunity to do so.

12.2.8. Averting Instrumental Incentives

Omohundro[117] notes that highly capable AI systems should be expected to pursue certain convergent instrumental strategies, such as the preservation of the system's current goals and the acquisition of resources. Omohundro's argument is that most objectives imply that an agent pursuing the objective should (1) ensure that nobody redirects the agent toward different objectives, as then the current objective would not be achieved; (2) ensure that the agent is not destroyed, as then the current objective would not be achieved; (3) become more resource-efficient; (4) acquire more resources, such as computing resources and energy sources; and (5) improve cognitive capacity.

It is difficult to define practical objective functions that resist these pressures.[118] For example, if the system is rewarded for shutting down when the humans want it to shut down, then the system has incentives to take actions that make the humans want to shut it down.[119]

A number of value-learning proposals, such as those discussed by Hadfield-Menell et al.,[120] describe systems that would avert instrumental incentives by dint of their uncertainty about which goal they are supposed to optimize. A system that believes that the operators (and only the operators) possess knowledge of the "right" objective function might be very careful in how it deals with the operators, and this caution could counteract potentially harmful default incentives.

This, however, is not the same as *eliminating* the relevant set of incentives. If a value-learning system were ever confidently wrong, the standard instrumental incentives would reappear immediately. For instance, if the value learning framework were set up slightly incorrectly, and the system gained high confidence that humans terminally value the internal sensation of pleasure, it might acquire strong incentives to acquire a large amount of resources that it could use to put as many humans as possible on opiates.

If we could design objective functions that averted these default incentives outright, that would be a large step toward answering the concerns raised by Bostrom[121] and others, many of which stem from the fact that these subgoals naturally arise from almost any goal.

12.2.8.1. Related Work

Soares et al.[122] and Orseau and Armstrong[123] discuss specific designs that can avert specific instrumental incentives, such as the incentive to manipulate a

shutdown button or the incentive to avoid being interrupted. However, these approaches have major shortcomings (discussed in those papers), and a satisfactory solution will require more research.

Those authors pursue methods for averting specific instrumental pressures—namely, pressure to avoid being shut down. However, it may be that there is a general solution to problems of this form, which can be used to simultaneously avert numerous instrumental pressures (including, e.g., the incentive to outsource computation to the environment). A general-purpose method for averting all instrumental pressures like the ones Omohundro describes—both foreseen and unforeseen—would make it significantly easier to justify confidence that an AI system will behave in a robustly beneficial manner. As such, this topic of research seems well worth pursuing.

12.2.8.2. Directions for Future Research

Soares et al.,[124] Armstrong,[125] and Orseau and Armstrong[126] study methods for combining objective functions in such a way that human operators have the ability to switch which function an agent is optimizing, but the agent does not have incentives to cause or prevent this switch. All three approaches leave much to be desired, and further research along those paths seems likely to be fruitful.

In particular, we would like a way of switching between objective functions such that the AI system (1) has an incentive to protect its operators' ability to switch between its objective functions; (2) has no incentive to try to control which objective function its operators switch to; and (3) has reasonable beliefs about the relation between its actions and the mechanism that switches its objective function. We do not yet know of a solution that satisfies all of these desiderata. Perhaps a solution to this problem will generalize to also allow the creation of an AI system that has no incentive to change, for example, the amount of computational resources it has access to.

Another approach is to consider creating systems that "know they are flawed" in some sense. The idea would be that the system would want to shut down as soon as it realizes that its operators are attempting to shut it down because it believes that its operators' judgment is less "flawed" than its own. It is difficult to formalize such an idea; naive attempts result in a system that attempts to model the ways it could be flawed and optimize according to a mixture over all of its different possible flaws, which is problematic if the model of various possible flaws is itself flawed. While it is not at all clear how to make this desired kind of reasoning more concrete, success at formalizing it could result in entirely new approaches to the problem of averting instrumental incentives.

12.3. Summary

A better understanding of any of the eight open research areas we have described would improve our ability to design robust and reliable AI systems in the future. To review:

1, 2, 3—A better understanding of robust inductive ambiguity identification, human imitation, and informed oversight would aid in the design of systems that can be safely overseen by human operators (and which query the humans when necessary).
4—Better methods for specifying environmental goals would make it easier to design systems that are pursuing the objectives that we actually care about.
5, 6, 7—A better understanding of conservative concepts, low-impact measures, and mild optimization would make it easier to design advanced AI systems that fail gracefully and admit of online testing and modification.
8—A general-purpose strategy for averting convergent instrumental subgoals would help us build systems that lack any incentive to deceive their operators, compete for resources, or otherwise behave in an adversarial fashion.

In working on problems like those discussed in this chapter, it is important to keep in mind that they are intended to address whatever long-term concerns with highly intelligent systems we can predict in advance. Solutions that would work for modern systems but would predictably fail for highly capable systems are unsatisfactory, as are solutions that work in theory but are prohibitively expensive in practice.

These eight areas of research demonstrate that there are open technical problems—some of which are already receiving a measure of academic attention—whose investigation is likely to be helpful down the road for practitioners attempting to actually build robustly beneficial advanced ML systems.[127]

Notes

1. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).
2. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (New York: Oxford University Press, 2014).
3. For examples, refer to Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong, "Corrigibility," paper presented at 1st International Workshop on AI

and Ethics at AAAI-2015, Austin, TX, 2015; Stuart Armstrong, "Motivated Value Selection for Artificial Agents," paper presented at 1st International Workshop on AI and Ethics at AAAI-2015, Austin, TX, 2015.

4. Stuart J. Russell, "Of Myths and Moonshine," *Edge*, November 14, 2014, http://edge.org/conversation/the-myth-of-ai#26015.
5. Nate Soares and Benja Fallenstein, "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda," in *The Technological Singularity: Managing the Journey*, ed. Victor Callaghan, Jim Miller, Roman Yampolskiy, and Stuart Armstrong, The Frontiers Collection (Springer, 2017).
6. Paul Christiano, "Scalable AI Control," *AI Control*, December 5, 2015, https://medium.com/ai-control/scalable-ai-control-7db2436feee7.
7. Lihong Li, Michael L. Littman, and Thomas J. Walsh, "Knows What It Knows: A Framework for Self-Aware Learning," In *25th International Conference on Machine Learning* (Helsinki, Finland: ACM, 2008), 568–75.
8. Kelvin Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." 2015.
9. Volodymyr Mnih et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 518, no. 7540 (2016): 529–33.
10. David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016): 484–503.
11. Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum, "Human-Level Concept Learning through Probabilistic Program Induction," *Science* 350, no. 6266 (2015): 1332–38.
12. As per Volodymyr Mnih et al., "Playing Atari with Deep Reinforcement Learning," paper presented at Deep Learning Workshop at Neural Information Processing Systems 26, Lake Tahoe, NV, 2013.
13. For a simple example of this sort of behavior with a fairly weak reinforcement learner, refer to Tom Murphy, "The First Level of Super Mario Bros. Is Easy with Lexicographic Orderings and Time Travel," *SIGBOVIK* (2013): 112–33.
14. This scenario is discussed in detail in Bostrom, *Superintelligence*, and Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (New York: Oxford University Press, 2008), 308–45. Tsvi Benson-Tilsen and Nate Soares, "Formalizing Convergent Instrumental Goals," paper presented at 2nd International Workshop on AI, Ethics and Society at AAAI-2016, Phoenix, AZ, 2016, provide a simple formal illustration.
15. Soares and Fallenstein, "Agent Foundations for Aligning Machine Intelligence with Human Interests."
16. Dario Amodei et al., "Concrete Problems in AI Safety," (2016).
17. Ibid., 20.
18. Stuart J. Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter," *AI Magazine* 36, no. 4 (2015).
19. Nate Soares, "The Value Learning Problem," paper presented at Ethics for Artificial Intelligence Workshop at IJCAI-16, New York, 2016.

20. Hubert L. Dreyfus and Stuart E. Dreyfus, "What Artificial Experts Can and Cannot Do," *AI & Society* 6, no. 1 (1992): 18–26.
21. Tom Dietterich, personal conversation, 2016.
22. Amodei et al., "Concrete Problems in AI Safety."
23. Soares and Fallenstein, "Agent Foundations for Aligning Machine Intelligence with Human Interests."
24. As an example of where our categorization of these problems differs: the concrete problems agenda considers "scalable oversight" to be a separate problem from robustness to distributional change. We instead place the problem of identifying situations where the training data are insufficient to specify the correct reward function under the umbrella of inductive ambiguity identification.
25. Alexander Genkin, David D. Lewis, and David Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics* 49, no. 3 (2007): 291–304; Charles Blundell et al., "Weight Uncertainty in Neural Networks," 2015; Anoop Korattikara et al., "Bayesian Dark Knowledge," 2015.
26. Huan Liu and Hiroshi Motoda, *Computational Methods of Feature Selection* (CRC Press, 2007); Yuhong Guo and Dale Schuurmans, "Convex Structure Learning for Bayesian Networks: Polynomial Feature Selection and Approximate Ordering," 2012.
27. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," 2014; Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), 427–36.
28. Yarin Gal and Zoubin Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," In *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)* (New York, NY: ACM, 2016), 353–60.
29. Burr Settles, "Active Learning Literature Survey," (Wisconsin, Madison: University of Wisconsin). https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf.
30. H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky, "Query by Committee," in *5th Annual Workshop on Computational Learning Theory* (ACM, 1992), 287–94; Alina Beygelzimer, Sanjoy Dasgupta, and John Langford, "Importance Weighted Active Learning," in *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, Quebec, Canada: ACM, 2009), 49–56.
31. Steve Hanneke, "A Bound on the Label Complexity of Agnostic Active Learning," in *Proceedings of the 24th International Conference on Machine Learning* (ACM, 2007), 353–60.
32. Steve Hanneke, "Theory of Disagreement-Based Active Learning," *Foundations and Trends in Machine Learning* 7, nos. 2–3 (2014): 131–309.
33. Alina Beygelzimer et al., "Search Improves Label for Active Learning," 2016.
34. Some forms of online active learning (refer to, e.g., Ofer Dekel, Claudio Gentile, and Karthik Sridharan, "Selective Sampling and Active Learning from Single and Multiple Teachers," *Journal of Machine Learning Research* 13, no. 1 [2012]: 2655–97)

relax the i.i.d. assumption, but the authors do not see how to apply them to the problem of inductive ambiguity identification.

35. See, for example, Ling Huang et al., "Adversarial Machine Learning," in *4th ACM Workshop on Security and Artificial Intelligence* (Chicago, IL: ACM, 2011), 43–58.
36. Vladimir Vovk, Alex Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World* (Springer Science & Business Media, 2005).
37. Li, Littman, and Walsh, "Knows What It Knows,"
38. Ibid.; Khani and Rinard 2016).
39. István Szita and Csaba Szepesvári, "Agnostic KWIK Learning and Efficient Approximate Reinforcement Learning." *JMLR: Workshop and Conference Proceedings* 19 (2011): 739–72.
40. Paul Christiano, "Active Learning for Opaque, Powerful Predictors," *Medium*, January 3, 2016, https://medium.com/ai-control/active-learning-for-opaque-powerful-predictors-94724b3adf06.
41. Paul Christiano, "Mimicry and Meeting Halfway," *Medium*, September 19, 2015, https://medium.com/ai-control/mimicry-maximization-and-meeting-halfway-c149dd23fc17.
42. Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25, no. 2 (2012): 1097–105; Kaiming He et al., "Deep Residual Learning for Image Recognition," 2015.
43. Andrej Karpathy and Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, June, 2015.
44. As studied by, for example, Karol Gregor et al., "DRAW: A Recurrent Neural Network for Image Generation," 2015; Lake, Salakhutdinov, and Tenenbaum, "Human-Level Concept Learning through Probabilistic Program Induction."
45. For example, Kshitij Judah et al., "Active Imitation Learning: Formal and Practical Reductions to I.I.D. Learning," *Journal of Machine Learning Research* 15 (2014): 4105–43; Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell, "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning," 2010; Tamim Asfour et al., "Imitation Learning of Dual-Arm Manipulation Tasks in Humanoid Robots," *International Journal of Humanoid Robotics* 5, no. 2 (2008): 183–202.
46. Andrew Y. Ng and Stuart J. Russell, "Algorithms for Inverse Reinforcement Learning," in *17th International Conference on Machine Learning (ICML-'00)*, ed. Pat Langley (San Francisco: Morgan Kaufmann, 2000), 663–70.
47. Pieter Abbeel and Andrew Y. Ng, "Apprenticeship Learning via Inverse Reinforcement Learning," in *21st International Conference on Machine Learning (ICML-'04)* (Ban, AB, Canada: ACM, 2004). http://doi.acm.org/10.1145/1015330.1015430.
48. Ziebart et al. (2008)
49. Pieter Abbeel, Adam Coates, and Andrew Ng, "Autonomous Helicopter Aerobatics through Apprenticeship Learning," *International Journal of Robotics Research* (2010).

50. Chelsea Finn, Sergey Levine, and Pieter Abbeel, "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization," 2016.
51. Owain Evans, Andreas Stuhlmüller, and Noah Goodman, *Learning the Preferences of Ignorant, Inconsistent Agents,* ed. Dale Schuurmans and Michael Wellman (Menlo Park, CA: AAAI Press, 2016); Owain Evans, Andreas Stuhlmüller, and Noah Goodman, "Learning the Preferences of Bounded Agents," 6, 2015. https://www.fhi.ox.ac.uk/wp-content/uploads/nips-workshop-2015-website.pdf.
52. Ian J. Goodfellow et al., "Generative Adversarial Networks," 2014.
53. Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes," 2013.
54. Andreas Stuhlmüller, Jessica Taylor, and Noah Goodman, "Learning Stochastic Inverses," in *Advances in Neural Information Processing Systems* (2013), 3048–56.
55. Paul Christiano, "Abstract Approval-Direction," *AI Control*, November 28, 2015, https://medium.com/ai-control/abstract-approval-direction-dc5a3864c092; Paul Christiano, "Approval-Directed Algorithm Learning," *AI Control*, February 21, 2016, https://medium.com/ai-control/approval-directed-algorithm-learning-bf1f8fad42cd.
56. W. Bradley Knox and Peter Stone, "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework," in *Proceedings of the Fifth International Conference on Knowledge Capture* (ACM, 2009), 9–16.
57. Paul Christiano, "The Informed Oversight Problem," *Medium*, April 1, 2016, https://medium.com/ai-control/the-informed-oversight-problem-1b51b4f66b35.
58. The lowest-order bits of the image would be uniformly random if the generator system were generating pictures as intended.
59. Knox and Stone, "Interactively Shaping Agents via Human Reinforcement."
60. Christiano, "The Informed Oversight Problem."
61. Christian Daniel et al., "Active Reward Learning," in *Proceedings of Robotics Science and Systems* (2014).
62. Dylan Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning," 2016.
63. Tom Everitt and Marcus Hutter, "Avoiding Wireheading with Value Reinforcement Learning," 2016.
64. Soares et al., "Corrigibility."
65. Amodei et al., "Concrete Problems in AI Safety."
66. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2013; Matthew D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision* (Springer, 2014), 818–33; Aravindh Mahendran and Andrea Vedaldi, "Understanding Deep Image Representations by Inverting Them," in *2015 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), 5188–96; Goodfellow, Shlens, and Szegedy, "Explaining and Harnessing Adversarial Examples."
67. Luca Pulina and Armando Tacchella, "An Abstraction-Refinement Approach to Verification of Artificial Neural Networks," in *International Conference on Computer Aided Verification* (Springer, 2010), 243–57.

68. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," 2016.
69. David Baehrens et al., "How to Explain Individual Classification Decisions," *Journal of Machine Learning Research* 11 (June 2010): 1803–31.
70. Anupam Datta, Shayak Sen, and Yair Zick, "Algorithmic Transparency via Quantitative Input Influence," in *Proceedings of 37th IEEE Symposium on Security and Privacy* (2016); Erik Štrumbelj and Igor Kononenko, "Explaining Prediction Models and Individual Predictions with Feature Contributions," *Knowledge and Information Systems* 41, no. 3 (2014): 647–65; Marko Robnik-Šikonja and Igor Kononenko, "Explaining Classifications for Individual Instances," *IEEE Transactions on Knowledge and Data Engineering* 20, no. 5 (2008): 589–600.
71. Alfredo Vellido, José David Martín-Guerrero, and Paulo Lisboa, "Making Machine Learning Models Interpretable," *ESANN* 12 (2012): 163–72.
72. Nir Friedman, Dan Geiger, and Moises Goldszmidt, "Bayesian Network Classifiers," *Machine Learning* 29, nos. 2–3 (1997): 131–63; Judea Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (New York: Cambridge University Press, 2009).
73. Philippe Weber et al., "Overview on Bayesian Networks Applications for Dependability, Risk Analysis and Maintenance Areas," *Engineering Applications of Artificial Intelligence* 25, no. 4 (2012): 671–82.
74. Iván Sánchez Carmona and Sebastian Riedel, "Extracting Interpretable Models from Matrix Factorization Models," in *COCO'15: Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches* 1583 (2015): 78–84.
75. Dominik Janzing et al., "Quantifying Causal Influences," *Annals of Statistics* 41, no. 5 (2013): 2324–58.
76. Benjamin Letham et al., "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model," *Annals of Applied Statistics* 9, no. 3 (2015): 1350–71.
77. Kim Baraka, Ana Paiva, and Manuela Veloso, "Expressive Lights for Revealing Mobile Service Robot State," paper presented at Robot'2015, the 2nd Iberian Robotics Conference, Lisbon, Portugal, 2015; Stephanie Rosenthal, Sai P. Selvaraj, and Manuela Veloso, "Verbalization: Narration of Autonomous Mobile Robot Experience," paper presented at 26th International Joint Conference on Artificial Intelligence, New York City, NY, 2016.
78. Andrea L. Thomaz and Cynthia Breazeal, "Transparency and Socially Guided Machine Learning," paper presented at 5th International Conference on Development and Learning, 2006.
79. Guangliang Li et al., "Using Informative Behavior to Increase Engagement While Learning from Human Reward," *Autonomous Agents and Multi-Agent Systems* 30, no. 5 (2015): 826–48.
80. Christiano, "The Informed Oversight Problem."
81. This practice of referring to preferences over world-states as utility functions dates back to John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press, 1944).

82. Daniel Dewey, "Learning What to Value," in *Artificial General Intelligence: 4th International Conference, AGI 2011*, ed. Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, Lecture Notes in Computer Science 6830 (Berlin: Springer, 2011), 309–14; Bill Hibbard, "Model-Based Utility Functions," *Journal of Artificial General Intelligence* 3, no. 1 (2012): 1–24.
83. Marcus Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*, Texts in Theoretical Computer Science (Berlin: Springer, 2005).
84. Everitt and Hutter, "Avoiding Wireheading with Value Reinforcement Learning."
85. Dewey, "Learning What to Value"; Amodei et al., "Concrete Problems in AI Safety."
86. Nicolas Heess et al., "Learning Continuous Control Policies by Stochastic Value Gradients," in *Advances in Neural Information Processing Systems 28*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, 2015), 2944–52.
87. This proposal is related to the work of Abel et al., who use a state-collapsing function ϕ for reinforcement-learning tasks with high-dimensional $\mathcal{S}$. Their agent explores by taking actions in state A that it hasn't yet taken in previous states B with $\phi(B) = \phi(A)$, where ϕ maps states to a small set of clusters. They achieve impressive results, suggesting that state-collapsing functions—perhaps mapping to a richer but still low-dimensional representation space—may capture the important structure of a reinforcement-learning task in a way that allows the agent to compare states to the goal state in a meaningful way. See David Abel et al., "Exploratory Gradient Boosting for Reinforcement Learning in Complex Domains," paper presented at *Abstraction in Reinforcement Learning Workshop at ICML-'16,* New York, 2016.
88. Mark Ring and Laurent Orseau, "Delusion, Survival, and Intelligent Agents," in *Artificial General Intelligence: 4th International Conference, (AGI 2011)*, ed. Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks (Berlin: Springer, 2011), 11–20.
89. This is the problem of "wireheading," studied by, among others, ibid.
90. Laurenz Wiskott and Terrence J. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances," *Neural Computation* 14, no. 4 (2002): 715–70.
91. Russell, "Of Myths and Moonshine."
92. Bostrom, *Superintelligence*.
93. Russell, "Of Myths and Moonshine."
94. In the simplest case, we can assume that these objects are specified as detailed 3D scans. If we have only incomplete observations of these objects, problems described in section 12.2.4 arise.
95. Goodfellow, Shlens, and Szegedy, "Explaining and Harnessing Adversarial Examples."
96. Ng and Russell, "Algorithms for Inverse Reinforcement Learning."
97. Goodfellow et al., "Generative Adversarial Networks."
98. Md Amran Siddiqui et al., "Finite Sample Complexity of Rare Pattern Anomaly Detection," in *Uncertainty in Artificial Intelligence: Proceedings of the 32nd*

Conference (UAI-2016), ed. Alexander Ihler and Dominik Janzing (Corvallis, OR: AUAI Press, 2016), 686–95.

99. Geoffrey E. Hinton and Ruslan R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science* 313, no. 5786 (2006): 504–7.
100. Amodei et al., "Concrete Problems in AI Safety."
101. Such as penalizing empowerment, as formalized by Christoph Salge, Cornelius Glackin, and Daniel Polani, "Empowerment: An Introduction," in *Guided Self-Organization: Inception* (Springer, 2014), 67–114.
102. Stuart Armstrong and Benjamin Levinstein, "Low Impact Artificial Intelligences," arXiv: 1705.10720 [cs.AI].
103. Armstrong suggests having those features be hand-selected, but they could plausibly also be generated from the system's own world-model.
104. Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision* 40, no. 2 (2000): 99–121.
105. Heess et al., "Learning Continuous Control Policies by Stochastic Value Gradients."
106. As formalized by Judea Pearl, *Causality: Models, Reasoning, and Inference* (New York: Cambridge University Press, 2000).
107. Russell, "Of Myths and Moonshine."
108. Amir M. Farahmand et al., "Regularized Policy Iteration," in *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, ed. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Curran Associates, 2009), 441–48.
109. Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds and Machines* 17, no. 4 (2007): 391–444.
110. Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv, "Empowerment: A Universal Agent-Centric Measure of Control," in *Evolutionary Computation, 2005* (IEEE, 2005), 1:128–35; Salge, Glackin, and Polani, "Empowerment."
111. For a discussion of this phenomenon, refer to Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto, "On Early Stopping in Gradient Descent Learning," *Constructive Approximation* 26, no. 2 (2007): 289–315; Geoffrey Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine* 29, no. 6 (2012): 82–97.
112. Steve Rabin, *Introduction to Game Development* (Nelson Education, 2010).
113. Bostrom, *Superintelligence*.
114. Herbert A. Simon, "Rational Choice and the Structure of the Environment," *Psychological Review* 63, no. 2 (1956): 129.
115. Jessica Taylor, "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization," paper presented at 2nd International Workshop on AI, Ethics and Society at AAAI-2016, Phoenix, AZ, 2015.
116. Ibid.
117. Stephen M. Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008: 1st AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin, Frontiers in Artificial Intelligence and Applications 171 (Amsterdam: IOS, 2008), 483–92.

118. Benson-Tilsen and Soares, "Formalizing Convergent Instrumental Goals."
119. Stuart Armstrong, "Motivated Value Selection for Artificial Agents."
120. Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning"; Soares, "The Value Learning Problem."
121. Bostrom, *Superintelligence*.
122. Soares et al., "Corrigibility."
123. Laurent Orseau and Stuart Armstrong, "Safely Interruptible Agents," in *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016)*, ed. Alexander Ihler and Dominik Janzing (Jersey City, NJ: 2016), 557–66.
124. Soares et al., "Corrigibility."
125. Armstrong, "Motivated Value Selection for Artificial Agents."
126. Orseau and Armstrong, "Safely Interruptible Agents."
127. Thanks to Paul Christiano for seeding many of the initial ideas for these research directions (and, to a lesser extent, Dario Amodei and Chris Olah). In particular, the problems of informed oversight and robust human imitation were both strongly influenced by Paul. Thanks to Nate Soares and Tsvi Benson-Tilsen for assisting in the presentation of this paper. Thanks to Stuart Armstrong for valuable discussion about these research questions, especially the problem of averting instrumental incentives. Thanks also to Jan Leike, Owain Evans, Stuart Armstrong, and Jacob Steinhardt for valuable conversations.

References

Abbeel, Pieter, and Andrew Y. Ng. "Apprenticeship Learning via Inverse Reinforcement Learning." In *21st International Conference on Machine Learning (ICML-'04)*. Ban, AB, Canada: ACM. http://doi.acm.org/10.1145/1015330.1015430.

Abbeel, Pieter, Adam Coates, and Andrew Ng. "Autonomous Helicopter Aerobatics through Apprenticeship Learning." *International Journal of Robotics Research* (2010).

Abel, David, Alekh Agarwal, Akshay Krishnamurthy Fernando Diaz, and Robert E. Schapire. "Exploratory Gradient Boosting for Reinforcement Learning in Complex Domains." Paper presented at Abstraction in Reinforcement Learning Workshop at ICML-'16. New York, 2016.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." 2016. arXiv: 1606.06565 [cs.AI].

Armstrong, Stuart. "Motivated Value Selection for Artificial Agents." Paper presented at 1st International Workshop on AI and Ethics at AAAI-2015. Austin, TX, 2015.

Armstrong, Stuart, and Benjamin Levinstein. "Low Impact Artificial Intelligences." arXiv: 1705.10720 [cs.AI].

Asfour, Tamim, Pedram Azad, Florian Gyarfas, and Rüdiger Dillmann. "Imitation Learning of Dual-Arm Manipulation Tasks in Humanoid Robots." *International Journal of Humanoid Robotics* 5, no. 2 (2008): 183–202.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. "How to Explain Individual Classification Decisions." *Journal of Machine Learning Research* 11 (June 2010): 1803–31.

Baraka, Kim, Ana Paiva, and Manuela Veloso. "Expressive Lights for Revealing Mobile Service Robot State." Paper presented at Robot'2015, the 2nd Iberian Robotics Conference. Lisbon, Portugal. 2015.

Benson-Tilsen, Tsvi, and Nate Soares. "Formalizing Convergent Instrumental Goals." Paper presented at 2nd International Workshop on AI, Ethics and Society at AAAI-2016, 62–70. Phoenix, AZ. 2016.

Beygelzimer, Alina, Sanjoy Dasgupta, and John Langford. "Importance Weighted Active Learning." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 49–56. ICML '09. Montreal, Quebec, Canada: ACM, 2009. 978-1-60558-516-1. doi:10.1145/1553374.1553381. http://doi.acm.org/10.1145/1553374.1553381.

Beygelzimer, Alina, Daniel Hsu, John Langford, and Chicheng Zhang. "Search Improves Label for Active Learning." 2016. arXiv: 1602.07265 [cs.LG].

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight Uncertainty in Neural Networks." 2015. arXiv: 1505.05424 [stat.ML].

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. New York: Oxford University Press, 2014.

Carmona, Iván Sánchez, and Sebastian Riedel. "Extracting Interpretable Models from Matrix Factorization Models." In *COCO'15: Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches* 1583 (2015): 78–84.

Christiano, Paul. "Abstract Approval-Direction." *AI Control*, November 28, 2015. https://medium.com/ai-control/abstract-approval-direction-dc5a3864c092.

Christiano, Paul. "Active Learning for Opaque, Powerful Predictors." *Medium*, January 3, 2016. https://medium.com/ai-control/active-learning-for-opaque-powerful-predictors-94724b3adf06.

Christiano, Paul. "Approval-Directed Algorithm Learning." *AI Control*, February 21, 2016. https://medium.com/ai-control/approval-directed-algorithm-learning-bf1f8fad42cd.

Christiano, Paul. "The Informed Oversight Problem." *Medium*, April 1, 2016. https://medium.com/ai-control/the-informed-oversight-problem-1b51b4f66b35.

Christiano, Paul. "Mimicry and Meeting Halfway." *Medium*, September 19, 2015. https://medium.com/ai-control/mimicry-maximization-and-meeting-halfway-c149dd23fc17.

Christiano, Paul. "Scalable AI Control." *AI Control*, December 5, 2015. https://medium.com/ai-control/scalable-ai-control-7db2436feee7.

Daniel, Christian, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. "Active Reward Learning." In *Proceedings of Robotics Science and Systems*. 2014.

Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence." In *Proceedings of 37th IEEE Symposium on Security and Privacy*. 2016.

Dekel, Ofer, Claudio Gentile, and Karthik Sridharan. "Selective Sampling and Active Learning from Single and Multiple Teachers." *Journal of Machine Learning Research* 13, no. 1 (2012): 2655–97.

Dewey, Daniel. "Learning What to Value." In *Artificial General Intelligence: 4th International Conference, AGI 2011*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 309–14. Lecture Notes in Computer Science 6830. Berlin: Springer, 2011.

Dreyfus, Hubert L., and Stuart E. Dreyfus. "What Artificial Experts Can and Cannot Do." *AI & Society* 6, no. 1 (1992): 18–26.

Evans, Owain, Andreas Stuhlmüller, and Noah Goodman. "Learning the Preferences of Bounded Agents." 6 (2015). https://www.fhi.ox.ac.uk/wp-content/uploads/nips-workshop-2015-website.pdf.

Evans, Owain, Andreas Stuhlmüller, and Noah Goodman. *Learning the Preferences of Ignorant, Inconsistent Agents*. Edited by Dale Schuurmans and Michael Wellman. Menlo Park, CA: AAAI Press, 2016.

Everitt, Tom, and Marcus Hutter. "Avoiding Wireheading with Value Reinforcement Learning." 2016. arXiv: 1605.03143 [cs.AI].

Farahmand, Amir M., Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. 2009. "Regularized Policy Iteration." In *Advances in Neural Information Processing Systems 21 (NIPS 2008*), edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, 441–48. Curran Associates, 2009.

Finn, Chelsea, Sergey Levine, and Pieter Abbeel. "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization." 2016. arXiv: 1603.00448 [cs.LG].

Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian Network Classifiers." *Machine Learning* 29, nos. 2–3 (1997): 131–63.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 353–60. New York, NY: ACM, 2016.

Genkin, Alexander, David D. Lewis, and David Madigan. "Large-Scale Bayesian Logistic Regression for Text Categorization." *Technometrics* 49, no. 3 (2007): 291–304.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." 2014. arXiv: 1406.2661 [stat.ML].

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." 2014. arXiv: 1412.6572 [stat.ML].

Gregor, Karol, Ivo Danihelka, Alex Graves, and Daan Wierstra. "DRAW: A Recurrent Neural Network for Image Generation." 2015. arXiv: 1502.04623 [cs.CV].

Guo, Yuhong, and Dale Schuurmans. "Convex Structure Learning for Bayesian Networks: Polynomial Feature Selection and Approximate Ordering." 2012. arXiv: 1206.6832 [cs.LG].

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "Cooperative Inverse Reinforcement Learning." 2016. arXiv: 1606.03137 [cs.AI].

Hanneke, Steve. "A Bound on the Label Complexity of Agnostic Active Learning." In *Proceedings of the 24th International Conference on Machine Learning*, 353–60. ACM, 2007.

Hanneke, Steve. "Theory of Disagreement-Based Active Learning." *Foundations and Trends in Machine Learning* 7, nos. 2–3 (2014): 131–309.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." 2015. arXiv: 1512.03385 [cs.CV].

Heess, Nicolas, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. 2015. "Learning Continuous Control Policies by Stochastic Value Gradients." In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2944–52. Curran Associates, 2015.

Hibbard, Bill. "Model-Based Utility Functions." *Journal of Artificial General Intelligence* 3, no. 1 (2012): 1–24.

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313, no. 5786 (2006): 504–7.

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29, no. 6 (2012): 82–97.

Huang, Ling, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. "Adversarial Machine Learning." In *4th ACM Workshop on Security and Artificial Intelligence*, 43–58. Chicago, IL: ACM, 2011.

Hutter, Marcus. 2005. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Texts in Theoretical Computer Science. Berlin: Springer, 2005.

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, and others. "Quantifying Causal Influences." *Annals of Statistics* 41, no. 5 (2013): 2324–58.

Judah, Kshitij, Alan P. Fern, Thomas G. Dietterich, and Prasad Tadepalli. "Active Imitation Learning: Formal and Practical Reductions to I.I.D. Learning." *Journal of Machine Learning Research* 15 (2014): 4105–43.

Karpathy, Andrej, and Li Fei-Fei. "Deep Visual-Semantic Alignments for Generating Image Descriptions." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition. June, 2015.

Khani, Fereshte, and Martin Rinard, and Percy Liang. "Unanimous Prediction for 100% Precision with Application to Learning Semantic Mappings." *arXiv*, last revised June 23, 2016. arXiv:1606.06368.

Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." 2013. arXiv: 1312.6114 [cs.LG].

Klyubin, Alexander S., Daniel Polani, and Chrystopher L. Nehaniv. "Empowerment: A Universal Agent-Centric Measure of Control." In *Evolutionary Computation, 2005*, 1:128–35. IEEE, 2005.

Knox, W. Bradley, and Peter Stone. "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework." In *Proceedings of the Fifth International Conference on Knowledge Capture*, 9–16. ACM, 2009.

Korattikara, Anoop, Vivek Rathod, Kevin Murphy, and Max Welling. "Bayesian Dark Knowledge." 2015. arXiv: 1506.04416 [cs.LG].

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25, no. 2 (2012): 1097–105.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-Level Concept Learning through Probabilistic Program Induction." *Science* 350, no. 6266 (2015): 1332–38.

Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no. 4 (2007): 391–444.

Letham, Benjamin, Cynthia Rudin, Tyler McCormick, David Madigan, and others. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model." *Annals of Applied Statistics* 9, no. 3 (2015): 1350–71.

Li, Guangliang, Shimon Whiteson, W Bradley Knox, and Hayley Hung. "Using Informative Behavior to Increase Engagement While Learning from Human Reward." *Autonomous Agents and Multi-Agent Systems* 30, no. 5 (2015): 826–48.

Li, Lihong, Michael L. Littman, and Thomas J. Walsh. "Knows What It Knows: A Framework for Self-Aware Learning." In *25th International Conference on Machine Learning*, 568–75. Helsinki, Finland: ACM, 2008.

Liu, Huan, and Hiroshi Motoda. *Computational Methods of Feature Selection*. CRC Press, 2007.

Mahendran, Aravindh, and Andrea Vedaldi. "Understanding Deep Image Representations by Inverting Them." In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 5188–96. IEEE, 2015.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. "Playing Atari with Deep Reinforcement Learning." Paper presented at Deep Learning Workshop at Neural Information Processing Systems 26. Lake Tahoe, NV, 2013. arXiv: 1312.5602 [cs.LG].

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518, no. 7540 (2016): 529–33.

Murphy, Tom. "The First Level of Super Mario Bros. Is Easy with Lexicographic Orderings and Time Travel." *SIGBOVIK*, 112–33. 2013.

Neumann, John von, and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press, 1944.

Ng, Andrew Y., and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning." In *17th International Conference on Machine Learning (ICML-'00)*, edited by Pat Langley, 663–70. San Francisco: Morgan Kaufmann, 2000.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 427–36. IEEE, 2015.

Omohundro, Stephen M. "The Basic AI Drives." In *Artificial General Intelligence 2008: 1st AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–92. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, 2008.

Orseau, Laurent, and Stuart Armstrong. "Safely Interruptible Agents." In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016)*, edited by Alexander Ihler and Dominik Janzing, 557–66. Jersey City, NJ. 2016.

Orseau, Laurent, and Mark Ring. "Self-Modification and Mortality in Artificial Agents." In *Artificial General Intelligence: 4th International Conference, AGI 2011*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 1–10. Lecture Notes in Computer Science 6830. Berlin: Springer, 2011.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press, 2000.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press, 2009.

Pulina, Luca, and Armando Tacchella. "An Abstraction-Refinement Approach to Verification of Artificial Neural Networks." In *International Conference on Computer Aided Verification*, 243–57. Springer. 2010.

Rabin, Steve. *Introduction to Game Development*. Nelson Education, 2010.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier." 2016. arXiv: 1602.04938 [cs.LG].

Ring, Mark, and Laurent Orseau. "Delusion, Survival, and Intelligent Agents." In *Artificial General Intelligence: 4th International Conference, (AGI 2011)*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 11–20. Berlin: Springer, 2011.

Robnik-Šikonja, Marko, and Igor Kononenko. "Explaining Classifications for Individual Instances." *IEEE Transactions on Knowledge and Data Engineering* 20, no. 5 (2008): 589–600.

Rosenthal, Stephanie, Sai P. Selvaraj, and Manuela Veloso. "Verbalization: Narration of Autonomous Mobile Robot Experience." Paper presented at 26th International Joint Conference on Artificial Intelligence, 862–68. New York City, NY, 2016.

Ross, Stéphane, Geoffrey J. Gordon, and J. Andrew Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning." 2010. arXiv: 1011.0686 [cs.LG].

Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. "The Earth Mover's Distance as a Metric for Image Retrieval." *International Journal of Computer Vision* 40, no. 2 (2000): 99–121.

Russell, Stuart J. "Of Myths and Moonshine." *Edge*, November 14, 2014. http://edge.org/conversation/the-myth-of-ai#26015.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

Russell, Stuart J., Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter." *AI Magazine* 36, no. 4 (2015): 105–14.

Salge, Christoph, Cornelius Glackin, and Daniel Polani. "Empowerment: An Introduction." In *Guided Self-Organization: Inception*, 67–114. Springer. 2014.

Settles, Burr. "Active Learning Literature Survey." Wisconsin, Madison: University of Wisconsin. https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf.

Seung, H. Sebastian, Manfred Opper, and Haim Sompolinsky. "Query by Committee." In *5th Annual Workshop on Computational Learning Theory*, 287–94. ACM, 1992.

Siddiqui, Md Amran, Alan Fern, Thomas G. Dietterich, and Shubhomoy Das. "Finite Sample Complexity of Rare Pattern Anomaly Detection." In *Uncertainty in Artificial Intelligence: Proceedings of the 32nd Conference (UAI-2016)*, edited by Alexander Ihler and Dominik Janzing, 686–95. Corvallis, OR: AUAI Press, 2016.

Silver, David, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (2016): 484–503.

Simon, Herbert A. "Rational Choice and the Structure of the Environment." *Psychological Review* 63, no. 2 (1956): 129.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." 2013. arXiv: 1312.6034 [cs.CV].

Soares, Nate. "The Value Learning Problem." Paper presented at Ethics for Artificial Intelligence Workshop at IJCAI-16. New York, 2016.

Soares, Nate, and Benja Fallenstein. "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda." In *The Technological Singularity: Managing the Journey*, edited by Victor Callaghan, Jim Miller, Roman Yampolskiy, and Stuart Armstrong, 103–25. The Frontiers Collection. Springer, 2017.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. "Corrigibility." Paper presented at 1st International Workshop on AI and Ethics at AAAI-2015. Austin, TX, 2015.

Stuhlmüller, Andreas, Jessica Taylor, and Noah Goodman. "Learning Stochastic Inverses." In *Advances in Neural Information Processing Systems*, 3048–56. 2013.

Szita, István, and Csaba Szepesvári. "Agnostic KWIK Learning and Efficient Approximate Reinforcement Learning." *JMLR: Workshop and Conference Proceedings* 19 (2011): 739–72.

Štrumbelj, Erik, and Igor Kononenko. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41, no. 3 (2014): 647–65.

Taylor, Jessica. "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization." Paper presented at 2nd International Workshop on AI, Ethics and Society at AAAI-2016. Phoenix, AZ, 2015.

Thomaz, Andrea L., and Cynthia Breazeal. "Transparency and Socially Guided Machine Learning." Paper presented at 5th International Conference on Development and Learning. 2006.

Vellido, Alfredo, José David Martín-Guerrero, and Paulo Lisboa. "Making Machine Learning Models Interpretable." *ESANN* 12 (2012): 163–72.

Vovk, Vladimir, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

Weber, Philippe, Gabriela Medina-Oliva, Christophe Simon, and Benoit Iung. "Overview on Bayesian Networks Applications for Dependability, Risk Analysis and Maintenance Areas." *Engineering Applications of Artificial Intelligence* 25, no. 4 (2012): 671–82.

Wiskott, Laurenz, and Terrence J. Sejnowski. "Slow Feature Analysis: Unsupervised Learning of Invariances." *Neural Computation* 14, no. 4 (2002): 715–70.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." 2015. arXiv: 1502.03044 [cs.LG].

Yao, Yuan, Lorenzo Rosasco, and Andrea Caponnetto. "On Early Stopping in Gradient Descent Learning." *Constructive Approximation* 26, no. 2 (2007): 289–315.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–45. New York: Oxford University Press, 2008.

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In *European Conference on Computer Vision*, 818–33. Springer, 2014.

Ziebart, Brian D., Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. "Maximum Entropy Inverse Reinforcement Learning." *AAAI* (2008): 1433–38.

13

Moral Machines

From Value Alignment to Embodied Virtue

Wendell Wallach and Shannon Vallor

Implementing sensitivity to norms, laws, and human values in computational systems has transitioned from philosophical reflection to an actual engineering challenge. The "value alignment" approach is among those that have gained traction with AI researchers, a subset of whom are primarily concerned about the safety of advanced artificial intelligence, or superintelligence. The value-alignment strategy posits that values can be learned by observing human behavior. In its initial conception it discarded the languages of normative ethics in favor of more computationally friendly concepts, such as utility functions, system goals, agent preferences, and value optimizers. Yet unlike concepts of justice, benevolence, duty, and virtue, the conceptual tools of the value-alignment approach carry no intrinsic ethical significance. While many defenders of value alignment may see their approach as simply a practical translation of utilitarian ethics, that is, as a mechanical path to an ideally rational and ethical decision calculus by means of a machine-learning method for understanding human preferences, there remain significant conceptual slippages in these translations. In this chapter we consider what may be lost in the excision of intrinsically ethical concepts from the project of engineering moral machines. We argue here that human-level AI and superintelligent systems can be assured to be safe and beneficial only if they embody something like *virtue* or moral character. Virtue embodiment is a more appropriate long-term goal for AI safety research than value alignment.

13.1. Moral Machines and Value Alignment

Breakthroughs in the deep-learning approach to artificial intelligence (AI) have been accompanied by an expanded interest in the safety of increasingly sophisticated systems and in the values that will inform their choices and actions. AI safety itself is a largely new research trajectory within the field of AI. AI researchers had been more focused on functionality—how to get a system to

Wendell Wallach and Shannon Vallor, *Moral Machines* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). © Oxford University Press. DOI: 10.1093/oso/9780190905040.003.0014

function so as to achieve a specified task. But there had been at least theoretical consideration of the damage an advanced AI might wreak in single-minded pursuit of fulfilling its designated task.[1] With advances in machine learning (ML), and deep learning more specifically, this theoretical concern appeared more feasible even if it remained speculative and not imminent. Thus was born an emphasis on AI safety as a corrective to research focused solely upon the functionality of AI systems.

Within AI safety research, "value alignment" has been proposed by Stuart Russell and others as a means to ensure that the values embodied in the choices and actions of AI systems are in line with those of the people they serve.[2] Value alignment quickly caught on within the AI safety research community. Yet among AI researchers there was little appreciation that a research field already existed that, for more than a decade, had considered challenges inherent in assuring that the choices and actions of autonomous systems are safe and ethically or morally appropriate. This field has gone by many names, including machine morality, machine ethics, and computational ethics; its central topic is the theoretical and practical prospects for moral machines. This highly interdisciplinary field is largely made up of moral philosophers, computer scientists, legal theorists, and applied or practical ethicists. A dialogue between machine ethicists and AI safety researchers has been slow in starting but has more recently gained some momentum.

A core concern for many of the AI safety researchers attracted to value alignment is the need to ensure that any future artificial general intelligence (AGI) or superintelligence would be friendly to human values and aligned with human interests, survival, and needs. In contrast, those who identify with machine ethics have devoted more attention to ways in which nearer-term autonomous systems can be designed to assure appropriate behavior in relatively common situations. Of course, machine ethicists also consider challenges that will arise as increasingly sophisticated systems encounter ever more complex ethical dilemmas. Recently AI researchers working on value alignment have also begun to direct attention to ensuring that systems fulfill nearer-term tasks in an appropriate manner. Nevertheless we believe it fair to say that value alignment as a research trajectory is particularly focused upon laying foundations for an approach to values that can be scaled up to guarantee the safety and human-friendly behavior of AGI systems.

For many philosophers considering the prospect of imbuing computational systems with ethical behavior, machine ethics is a largely theoretical challenge. After all, moral philosophers and psychologists have yet to acquire a thorough understanding of *human* moral decision-making. Still, a few interdisciplinary teams have begun work on computational pathways for implementing moral decision-making capabilities in machine systems. However, the techniques

they utilize are not the machine-learning algorithms increasingly favored by AI researchers, but "top-down" methodologies of constraint by deontic moral logics, decision trees, and so on.[3]

Value-alignment researchers are clearly intent on avoiding the existential risks they believe are inevitable in the development of AGI. But the value-alignment project, as it was originally described, appeared hopelessly naive from the perspective of many moral philosophers and practical ethicists. First, "values" is a relatively nebulous term, perhaps selected as a means to avoid the more difficult issues entailed by ethics or morality. Second, observation of human behavior, from which value-alignment theorists aim to deduce the desired "values" to which machines should align their behavior, might reveal an individual's or a community's *preferences*, but it will not necessarily indicate what is right, good, just, or appropriate. For philosophers this is a failure to appreciate the is/ought distinction, or more broadly, the distinction between descriptive and normative ethical inquiry. The use of value and its entanglement with preferences misleadingly suggests that values can be reduced to observable facts, and that appropriate behavior can be reverse-engineered algorithmically. Yet moral philosophers will insist that these assumptions rest upon a conflation of moral and nonmoral concepts, and a failure to understand the moral concept of value as fundamentally *prescriptive*, that is, indicating what we *ought* to prefer, whether or not the facts of our own behavior obey this prescription. Even those moral philosophers who subscribe to *ethical naturalism* and thus reject the fact-value distinction will deny that moral facts are derivable simply from observed human preferences.

Defenders of the value alignment approach may grant the need to frame its goals and methods in more nuanced terms that acknowledge the complex distinctions between human preferences, behaviors, conventional norms, and ethical norms, and the challenge of building machines that can successfully distinguish and navigate them. Yet if "value alignment" is then simply taken to mean "whatever it takes to build safe and reliably ethical AI agents," then by definition it is the approach we need. However, this also empties the notion of any definite technical meaning, threatening to make the notion of value alignment benign but vacuous. Conversely, if "value alignment" designates the active technical project marked by particular methods such as inverse reinforcement learning, then it remains questionable whether and how that approach can reasonably hope to engineer AI systems capable of tracking and being steered by the richly textured, spontaneous, and constantly evolving fabric of human ethical life.

Scientists often feel that the issues raised by philosophers and practical ethicists make the determination of appropriate behavior more complicated than it needs to be. Within the engineering ethos, a process cannot be fully understood unless one tries to create or reproduce it. If this is correct, it would seem that to pursue an exclusively theoretical route to machine ethics gets things the

wrong way around; better to use technical means to try to *reproduce* moral action, and learn from our successes and failures along the way. Furthermore, as scientists and others often note, ethicists commonly differ in their judgments, and their approaches do not always lead to clear guidance on courses of action. Seemingly unresolvable moral dilemmas or "wicked" moral problems appear frequently in reflections within moral philosophy. Even applied ethicists acknowledge that there is often neither consensus nor a single optimal solution to many such moral challenges. The complexities inherent in the domain of social action are simply too great.

"Ethical decision-making cannot be reduced to an algorithm" has been asserted by many a moral philosopher; here the philosopher follows the counsel of Aristotle, who states in his *Nicomachean Ethics*, "It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits."[4] For our purposes the stress is on the last phrase. Aristotle goes on to argue, we think correctly, that the profound complexity and instability of human social and ethical life does not permit description or analysis of this domain to attain the same level of precision as we would rightly expect from careful description of mathematical objects and relations. But does this mean that ethics cannot offer precise and unambiguous action guidance? And if it cannot, then what good is it to AI research and design? Can the study of ethics provide any useful, practical insights to AI researchers seeking to build systems that are safe and controllable and whose actions can be guaranteed to be beneficial?

In this chapter we introduce some ideas and key concepts of moral philosophy that can be placed in the service of machine ethics and show how they can be applied to promote appropriate machine behavior from systems likely to be deployed over the next ten to fifteen years. However, we acknowledge that such approaches are unlikely to be sufficient to ensure ethical machine behavior when, or if, autonomous systems become capable of self-guided intelligent action across the full range of human contexts and settings. In uncontrolled and unrestricted settings, we argue, autonomous AI systems "in the wild"—up to and including AGI—are unlikely to become reliably safe and ethical actors in the absence of some machine analog to *embodied human virtue*. By "embodied human virtue" we mean the rootedness of moral excellence in the affective, perceptual, and habitual dimensions of the human body and its relationship to the environment it inhabits.

13.2. Core Concepts in Machine Ethics

A few basic distinctions have emerged for clarifying approaches to building moral machines. James Moor[5] distinguishes between machines that are implicit

ethical agents and those capable of making explicit moral decisions. Implicit ethical agents are those whose behavior has been constrained so they cannot perform ethically forbidden acts. Similarly, Allen, Smit, and Wallach[6] made a distinction between computational systems that are operationally moral, functionally moral, and artificial moral agents. Operationally moral systems are those that function within bounded moral contexts, in which the engineers and designers can discern in advance the array of challenges the machines will encounter. In effect, the computational system is programmed in advance to act appropriately in each situation it will encounter. To the extent that the behavior of these machines is imbued with values, they are the values of the designers and engineers who build the systems, or the values of the companies for whom they work. When designers and engineers cannot predetermine all the circumstances an artificial agent will encounter, it becomes necessary for the agent to have subroutines that facilitate making explicit moral decisions. Nevertheless, over the coming decade or two, most artificially intelligent agents will continue to be single-purpose machines operating in bounded moral contexts, and their explicit moral reasoning will be limited to determining which norms or courses of action apply in the situation at hand or when values conflict. For example, a caregiving robot attending to a homebound or elder person might have to select whether to deliver a meal or medicine on schedule or whether to stop and recharge its battery. The right course of action could depend on how critically the individual needs the specific medicine, what might occur if the agent fails to recharge its battery immediately, or other factors.

Given limitations in the cognitive capabilities of present-day AI systems, the contexts within which they can function appropriately are limited. However, as breakthroughs are made in machine learning, commonsense reasoning, planning, working with analogies, and language aptitude, the environments within which intelligent systems can operate safely and acceptably will expand.

Machine ethicists question how artificial moral agents will make appropriate choices as contexts get more complex, values conflict, and the systems have enough autonomy to encounter a broad array of ethically significant choices. Will they recognize the features of the context they are in, and therefore which norms or procedures apply? Can they prioritize values in a manner that will lead them to appropriate or acceptable actions, if not always the best course of action? Might they be able to make rudimentary analyses of the consequences of various responses to a challenge in order to pick one that appears to maximize welfare, or the "good" of those affected by their action? In which circumstances might they require additional cognitive capabilities, beyond being able to reason, in order to make good judgments? These capabilities might include emotional intelligence, a theory of mind, empathy, embodied intelligence, semantic understanding, and consciousness. The bottom line is practical. Will, and how might,

more sophisticated systems act appropriately or acceptably as their autonomy increases and they confront ever more complex contexts and situations?

Whether artificial agents will eventually be full moral agents capable of functioning autonomously in all situations and be worthy of rights and responsibility is an intriguing philosophical and legal question that goes well beyond the near-term practical challenges engineers will confront as they build single systems. Many in the AI community presume AGI and superintelligence are inevitable.[7] Their interest in AI safety and value alignment, as mentioned, is often driven by a desire to ensure that advanced AI either is controllable or will embody values that are sensitive to and protective of human needs. Whether focus on nearer-term ethical challenges will lay foundations for ensuring the value alignment of advanced systems or is largely irrelevant to meeting that more futuristic concern is a matter upon which thoughtful experts disagree. Of course, whether AGI or superintelligence is truly possible or likely to be realized in the next fifty to one hundred years is also a matter upon which experts both in and outside of the AI community disagree.

13.3. Values, Norms, Principles, and Procedures

Values and valuing pervade everything. A value can be grounded in a simple valence, such as a disposition to "like" or "dislike" something or someone, or in a subtle preference "for" some entity or state of affairs, regardless of whether that valence is rooted in an ethical concern, such as justice or benevolence. Values can also be understood as intrinsic and unconditional (e.g., the inextricable moral value of the life of a human person), or they can be seen as extrinsic (conditionally assigned to the valued entity by an external valuing agent). Within the neural networks favored by machine-learning researchers, values are commonly represented by valences connected to either a node or a collection of nodes that capture the characteristics of a percept. Simple Hebbian learning, the earliest of machine-learning techniques, can strengthen the connection between the nodes. These connections can also decay in strength over time if left unused. The difficulty lies in assuring that connectionist learning will actually capture the more nuanced and complex characteristics of ethical principles or procedures. That is, can complex values be represented computationally, and if so how? Or, as we will discuss in this chapter, can bottom-up learning be scaled so as to embody a virtuous character?

Because values can be so nebulous in their importance, meaning, and application, moral philosophers turn to other concepts and terms to represent higher-order or primary ethical concerns. These include norms, duties, principles, and procedures that inform judgments in morally significant situations. Norms refer

to standards, accepted practices, or proscribed behaviors. Within ethics, norms set standards as to the acceptability or permissibility of various forms of behavior. Norms are commonly context-specific; that is, the norm and/or its appropriate mode of expression can change as the context changes.[8] Thus the set of possible norm specifications is almost infinite. In theory, an AI system might learn or catalog all norms and the situations to which they apply. However, in practice this would imply a full recognition of the features of the context in which the artificial agent is embedded, in order to discern which norms apply and how they should be expressed. To complicate matters further, consider the fact that the introduction of an artificial agent into any social context will alter that context, adding yet another layer of computational complexity and uncertainty.

Higher-order principles that frame many approaches to ethics facilitate decision-making by introducing ethical goals or duties that are defined so broadly that they cover countless situations. For example, in bioethics broadly and medical ethics specifically, the duties of beneficence, nonmaleficence, respecting individual autonomy, and justice or fairness inform all ethical decision-making. Such principles might suggest a schema for algorithms that frame an agent's decisions in ways that aim to reduce contextual variability.

Higher-order principles also have weaknesses. The goals or duties they set are often defined so broadly and abstractly that specific applications are debatable. Static definitions of goals and duties can lead to situational inflexibility. Goals and duties can conflict, and a clear method for resolving such conflicts may not be available. Furthermore, top-down computational systems are commonly confronted with "framing" problems—problems in tracking the ethically salient features of a context or the ethical importance of a decision made in a complex environment.[9] The use of heuristics for solving such framing problems can be helpful, but may also compromise the integrity and clarity of principle-based reasoning.

Instead of seeking conformity to a multiplicity of principles, duties, or goals that may conflict, consequentialist ethics such as utilitarianism favors a procedural solution that maximizes a single goal, such as aggregate welfare or net happiness. In other words, the best course of action is not one where the agent follows the rules, duties, or principles, but rather one in which the agent determines which among the courses of action it might take will lead to the best outcome. Utilitarianism is particularly attractive to AI engineers. It appears to suggest that selecting the right course of action is, in principle, a straightforward exercise wherein the sum of undesirable consequences for each option is subtracted from the sum of desirable consequences, and the option with the largest positive value is the appropriate action. Just calculate! Furthermore, the utility-maximizing principle espoused by consequentialists appears to be similar to the utility functions that AI engineers are familiar with. The strength of

mathematical utility functions is that they can factor in a nearly infinite number of variables; that is, they can manage very difficult calculations. In practice, however, there are real differences between utilitarian calculations and what can be accomplished with an empirical utility function. First, there are definitional concerns. What is actually to be maximized? Is it net happiness? Is it net welfare? How are happiness and welfare defined, and what empirical measures will be used to calculate happiness or welfare?

More important, utilitarianism depends upon calculating the likely consequences even when it is difficult or impossible to know all the consequences that may result for each course of action or their respective probabilities. For example, how deep should the analysis go? Which secondary and tertiary consequences should be included? Is there a stopping procedure for limiting the depth of analysis? What about factoring in "normal accidents,"[10] low-probability events, or Black Swans[11]—unforeseen, low-probability, high-impact events? Simply put, we often lack adequate information to make satisfactory utilitarian determinations. This critique is commonly thrown at those espousing utilitarianism as a useful ethical theory. In defense, utilitarian theorists such as John Stuart Mill or the contemporary ethicist Peter Singer argue that it is, nevertheless, the "right" principle for distinguishing good actions from bad ones, and rough utilitarian determinations can be made.[12] In practice, those utilitarian decisions that are made factor in experience, intuition, and the capacity to imagine and plan possible courses of action and their outcomes. Imagination and planning are well beyond the cognitive capacities realizable in present-day AI systems. Whether future systems will have such capabilities is still unknown.

13.4. Top-Down, Bottom-Up, and Hybrid Approaches to Moral Machines

How helpful is ethical theory in building AI agents sensitive to value considerations and the factoring of these into their choices and actions? Scholars within the field of machine ethics have noted that ethical theory suggests two broad approaches to the design of the control architecture of moral machines: top-down and bottom-up.[13]

A top-down approach takes an antecedently specified ethical theory and analyzes its computational requirements to guide the design algorithms and subsystems capable of implementing the theory. For example, some machine ethicists have considered whether rules such as the Ten Commandments or Asimov's Laws of Robotics can be implemented computationally. Others have analyzed the computational requirements for instantiating Mill's utilitarianism,

Kant's categorical imperative, or the prima facie duties espoused by W. D. Ross, though none of these is without deep-seated problems of application and interpretation that resist algorithmic solution.[14]

While it is possible that children come into the world with an innate capacity for moral decision-making, they also generally learn what is acceptable or permissible and what is unacceptable from the bottom up, through experience and learning. If a bottom-up approach to designing a moral machine uses a prior theory at all, it does so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or a control structure. The strength of bottom-up systems lies in their ability to dynamically integrate inputs from discrete subsystems. One weakness is the difficulty in defining the goal a bottom-up system, such as a genetic algorithm, should be trying to actuate or maximize. Another difficulty entails assembling the many discrete components of an agent to operate as a functional whole.

Value alignment is a bottom-up approach. Both computational strategies that simulate evolution and machine learning suggest methods for designing algorithms that could facilitate bottom-up approaches for acquiring sensitivity to moral phenomena. However, the details of how the value-alignment problem will be solved through machine learning and evolutionary algorithms are unclear. Furthermore, the forms of machine learning presently available, even the rudimentary forms of unsupervised learning currently being explored, are not robust enough to simulate the structured and unstructured learning that facilitate a child's exploration of her relationships and environment in the acquisition of moral acumen. At this stage in the development of computational systems, we lack the tools for the kind of unstructured learning in which mental states, subtle emotional rewards, relationships with others, and punishment can play important roles.

Because neither top-down nor bottom-up approaches to machine ethics are likely to deliver the combination of contextual adaptivity and norm governance that full moral agents display, eventually we will need hybrid systems that integrate bottom-up learning with a capacity to subject the evaluation of choices and actions to top-down principles or procedures that represent ideals we strive to meet.[15] Such a system must maintain the dynamic morality of bottom-up approaches that accommodate diverse inputs. These include affective inputs that simulate the functional capabilities of moral sentiments and emotions that evolve from being embodied in a world with others and that inform capacities central to moral intellect, such as a theory of mind, social understanding, and sentience. Whether the mere simulation of such inputs will be sufficient, as opposed to their somatic and phenomenal instantiation, is unclear at this time.

13.5. The Limitations of a Hybrid Approach

We have seen that hybrid approaches to developing moral machines may be the most promising approach currently available, yet even these approaches will very likely fall short of supplying human-level moral intelligence. Unless AI is deployed by terrorists, it must be acknowledged that the floor of human moral behavior will remain well below that of machines (for even the least intelligent machine will not be actively malicious or determinedly evil). Our concern here is the *ceiling*—the comparison between the level of safety and moral security that the *best* people can offer us, and that which we can expect from our best moral machines. Even with hybrid approaches, we should expect moral machines to struggle in certain contexts involving moral choice, contexts that a morally intelligent and virtuous human agent would normally be capable of managing quite well. Such contexts include (1) contexts requiring the agent to reason creatively or to successfully negotiate and resolve "wicked" moral conflicts between competing values and duties; (2) contexts involving radically new situations or forms of moral choice for which existing rules, principles, and learned patterns of moral behavior are insufficient guidance; (3) contexts involving multiple stakeholders with very different motivations, goals, norms, and capacities, where the moral standing of each interested party must be discovered, or in some cases established, through cooperative and critical moral discourse; and (4) contexts in which the salient ethical features are novel and thus especially difficult to recognize or discern.

What all of these cases have in common is the need for a cluster of advanced moral capacities that even a hybrid approach to machine morality is likely to fall short in delivering. These include the following:

Creative moral reasoning—the ability to invent new and *appropriate* moral solutions in ways underdetermined by the past.

Moral discourse—the ability to identify, conceptually frame, and negotiate moral solutions through cooperative reasoning with other moral agents.

Critical moral reflection—the ability to stand back and critically evaluate one's own moral outlook, and that of others, from the moral point of view itself, that is, the capacity to form second-order normative evaluations of existing moral values, desires, rules, and reasons.

Moral discernment, which includes the capacity to recognize new or previously uncategorized forms of moral salience, as well as recognizing subtle moral tensions and conflicts that reveal unresolved ethical issues.

Holistic moral judgment—the ability to make sense of a complex situation in ways that transcend the sum of its composite ethical factors, with an eye toward actively constructing the best way to live, all things considered.

Again, it is uncontroversial that many if not most humans fail to cultivate these advanced moral capacities in themselves, or if they do, fail to deploy them consistently and well. But it is equally uncontroversial that *some* humans have cultivated these capacities and are able to deploy them, with varying degrees of practical success. The existence of human moral expertise, however fragile and rare, is a fact that not only informs but *sustains* the domain of ethics in human history. It is how ethical norms and standards are able to remain adaptive to changing social and physical environments. It is what makes ethics truly normative and open to progressive improvement rather than functioning merely as convention, or as "politics by other means." It is also what makes ethics our only reliable recourse for action guidance when the mechanisms of law, politics, or custom and convention fail or become corrupted in ways that endanger the well-being of the moral community.

Two common features mark these advanced moral capacities: their potential responsiveness to new or reconfigured moral phenomena and their support for holistic, qualitative judgments that "make sense" of the moral field as a whole, in ways that go beyond the addition and subtraction of explicit values embedded in its parts.[16] Machines that lack these advanced moral capacities will be incapable of managing the kinds of situations that require them, and if given unsupervised agency in those contexts, such "moral machines" may fail in ways that gravely endanger human interests. As long as we retain meaningful and robust human control of machine behavior, this need not preclude the responsible use of machines equipped with lesser degrees of moral capacity. After all, such machines may function well in the vast majority of ethical contexts, most of which are relatively mundane. They may even be *more consistently* successful in mundane contexts than will humans in aggregate, given our species' distinctive penchant for self-destructive, spiteful, unreasonably aggressive, and malicious conduct. Furthermore, as a species we are often distracted, inattentive, or neglectful of moral considerations, a trait unlikely to be passed on to moral machines. Thus if moral machines could be safely *confined* to these mundane settings, we might need to go no further than a bottom-up or hybrid approach that ensures close value alignment with whatever human moral conventions are operative in those settings.[17]

Yet such behavioral confinement by value-alignment strategies cannot be guaranteed, for two reasons: first, because a mundane moral context can easily be perturbed by a sudden change or development, one that causes an unpredictable spiral of an easily manageable situation into a "wicked" or unprecedented moral challenge that demands advanced moral competence. Second, it is a near certainty that the growing demand for, and expansions of, machine autonomy in a range of practical contexts will place increasing pressure on the safety mechanisms of human supervision and machine confinement.

13.6. Virtue Ethics and Virtuous Machines

What, then, must we do? What sort of machine could be trusted as an ethical agent *even* in those situations demanding advanced moral competence? We can find outlines of an answer in a normative account of *virtue ethics*: a type of approach to ethics that is grounded not in rules or consequences but in the distinctive character traits of morally excellent agents, traits such as practical wisdom, honesty, justice, and moderation.[18] Virtue ethics is frequently used as a model by those advocating hybrid and bottom-up theories of moral machine development,[19] but in those accounts virtue ethics is generally considered to be no more than an instructive pattern that may be helpful for AI researchers to imitate in various ways, rather than a standard of ethical agency that one aims to literally embody in a machine. Nevertheless, Wallach and Allen[20] note that machine virtues, if they could be embodied, would provide the kind of reliability in moral character we would need from more advanced artificial agents. This is because virtues function as context-adaptive skills that generally enable their possessors to navigate moral contexts successfully—even contexts that are novel or unusually challenging. While virtuous agents are not morally infallible, they reliably approximate the peak level of moral performance that can be asked or expected of trusted agents operating in a given social context.

There are good reasons, in fact, to think that virtue *cannot* be embodied in machines given the techniques and resources available to AI researchers today. We will articulate these reasons in what follows. Nevertheless the idea of a "virtuous machine," where this is understood as a literal attribution and not a loose, metaphorical allusion, can serve as a *regulative ideal* in machine morality and safety research. It can remind us of the gold standard of moral capacity that we must aim for if we ever hope to have machines that can be fully entrusted with our safety and well-being. It can also dictate a level of moral machine capacity *below which* we are duty-bound to ensure that our morally immature machines are properly supervised, constrained, and confined to whatever extent practically possible.

Virtue ethics—whether rooted in the extensive tradition associated with Aristotle's *Nicomachean Ethics*, in noneudaimonist and sentimentalist approaches such as Hume's, or in traditions of East Asia such as Confucianism and Buddhism that employ virtue-driven structures—concerns itself centrally with the character dispositions and refined practical wisdom of the moral agent.[21] It is less concerned with the moral rules the agent follows or the specific consequences that she brings about (although both are acknowledged to play a subordinate role in ethical life). By "character," a virtue ethicist means those robust habits and skills acquired by a person that produce a reliable behavioral disposition toward ethically appropriate action; such a reliable disposition to moral

excellence is called a *virtue*.[22] Examples of virtues are honesty, wisdom, courage, and benevolence. Those of poor moral character have acquired the opposite kind of behavioral dispositions; they are disposed toward ethically inappropriate and unsuccessful behavior; any such disposition is called a *vice*. Examples of vices are dishonesty, cowardice, foolishness, and cruelty.

Because the focus of virtue ethics is on character rather than isolated actions, a single act of truth-telling is not an instance of the virtue of honesty, for even a viciously dishonest person occasionally tells the truth. Moreover, sometimes the person with the virtue of honesty will lie, for virtue is always situationally appropriate and guided by what Aristotle called *phronesis* or practical wisdom—another term for creative moral intelligence. Virtue is never rigid, mindless, or marked by unreasonable conformity with a rule or convention. For example, an honest person will find it morally necessary, in certain contexts, to conceal the truth to avoid doing undue harm to another's feelings or reputation, and in other contexts will find it morally necessary to lie in order to honor a promise made to another to keep a secret. But these are not hard-and-fast rules either; there are other circumstances that would compel an honest person to tell the truth even at the cost of harm to another, and still additional contexts that would compel an honest person to disclose a secret she had previously promised to keep. This does *not* mean that virtue is subjective; in a particular context a certain act may be objectively morally impermissible for any virtuous agent. However, textbooks on ethical theory are rife with examples of "wicked" moral contexts involving conflicting duties, rules, and values, where any rigid normative guidance such as that offered by utilitarian or deontological frameworks seems to fail us.[23] This is in fact one of the chief advantages of virtue ethics: that it *accepts* the fluidity and unpredictability of moral life as a worldly fact, yet avoids ethical relativism or egoism by developing a theoretical account of the advanced moral capacities that allow virtuous persons to negotiate even wicked and novel moral problems in real time, with exemplary (though not infallible) success.

What defines moral "success" on this account? One might initially expect to find a ready agreement here between the virtue ethicist and the advocate of inverse reinforcement learning approaches to value alignment.[24] This is because an important standard of success in most virtue ethics frameworks is alignment of the agent with behavior modeled by moral exemplars in the community; in Aristotelian ethics, the *phronimoi,* or "practically wise" who inspire others to cultivate their virtues;[25] or in Confucian virtue ethics, the *junzi,* or "refined person."[26] Likewise in inverse reinforcement learning approaches to value alignment, the machine infers the appropriate values and behavioral constraints from its observations of human moral models acting in the relevant environment(s). Yet this parallel does not fully hold, for two reasons. The first has to do with the importance of the agent's internal state of moral cultivation, and the second has

to do with the more fundamental standard of moral success in virtue ethics, namely, objective social health and flourishing.

To unpack the first reason, we must note that virtue also requires the gradual acquisition of appropriate states of moral understanding, belief, and feeling that should accompany the agent's moral behaviors; without these aligned *internal* states, the agent's behavior may be prosocial but morally empty. If this was only a problem of the agent's "moral sincerity," it would not matter for machines; we needn't care if they understand and feel correctly, so long as they act correctly. We may want humans to do more than "moral pantomime," but for machines, consistently effective moral pantomime seems quite sufficient for our purposes. Unfortunately, the appropriate internal states serve another, more vital purpose. They ensure that the agent has acquired through moral modeling and practice the deeper understanding of the ethical field of human action that is indispensable in order to know when morality requires them to *modify, suspend, or otherwise deviate* from the behavioral pattern typically observed in the moral exemplar(s) they have learned from.[27]

The moral apprentice must eventually learn when to separate from the moral master, either because the master's pattern is not appropriate in a certain unique case, or because the master's pattern is no longer well adapted to new features of the social environment (e.g., in requiring more sustainable practices for resource use as populations grow), or because the master's pattern can be improved upon (e.g., by the moral movement away from patriarchal or other supremacist norms). This requires an understanding of moral salience and the associated ability to discriminate between morally relevant differences and morally irrelevant differences—something that may be a technically insurmountable challenge for machine ethics in the near term, and hard to imagine how to achieve over the longer term.

This is because in principle, anything can be morally relevant, and the set of circumstances in which a given thing could be morally relevant can never be enumerated or reduced to an explicit mathematical function. Whether a man is wearing white or blue socks would, for almost all cases that a machine or human apprentice can observe, make no difference to the treatment he receives from a virtuous person. A machine trained to align its behavior with human values through inverse reinforcement learning, or any other method, would almost certainly ignore the color of his socks, and rightly so. But it is not difficult to invent any number of scenarios, however rare, in which it would make an *enormous* moral difference what color the man's socks were (e.g., when the sock color was a preestablished signal of imminent danger or a gesture toward an intimate joke long shared between the two people). A reliably virtuous agent can make novel identifications of moral relevance within a practically infinite range of possible configurations of the social context; a machine would need the equivalent

background and capability in order to be a reliable moral agent "in the wild" of the human social fabric, where new variants and configurations of moral salience are continually emerging.

This isotropy problem of moral relevance has been articulated in discussions of the challenge of building autonomous lethal robots for military use, and it has been suggested by Guarini and Bello[28] that some sort of emotional processing may be needed for humans (and, by extension, machines) to handle the computational load it presents. So how *do* virtuous humans manage to discern when a novel form of moral salience has unexpectedly changed the appropriate pattern of moral response? The only satisfactory answer is that virtuous agents understand, in a holistic, integrated, and richly embodied sense, the fabric of moral life and are thus perceptually and somatically attuned to the constant perturbations of its edges, threads, and foldings. Building this understanding into a machine is perhaps the most opaque computational task we could imagine, other than tasks that would violate the laws of physics. Yet without it, machines trained to align their values with even the *best* existing models of ethical behavior will, in uncontrolled social environments, inevitably fail to act in ethical ways—potentially with grave consequences.

We said there was a second reason why a moral modeling approach to value alignment will not guarantee the reliable inculcation of machine virtue. This is because human exemplars of virtue do not themselves set the standard of virtuous behavior; they are merely the nearest embodiment of it that a moral apprentice can access. What is the true moral standard of virtue, then? Virtues are distinguished from vices by their tendency to promote *human flourishing* (a translation of the Greek term *eudaimonia*) over the long term. Human flourishing is not a mental state but a way of living in community, an objective condition of social health enjoyed by the individual in concert with other members of their shared environment.[29] Thus it is distinct from subjective happiness (though it tends to foster the latter more reliably than other ways of living), and is also distinct from passive conformity to social conventions, which may often fail to foster social health.

Now, there are a vast number of possible configurations of "social health," and so this view does not assume that there is one objectively best way to live that virtuous agents must track. Within the shifting possibility space created by the many biological, ecological, and technological constraints that define social life for the human animal, different cultural configurations of the good life are constantly being constructed and negotiated. However, certain sociocultural configurations can and do violate those objective constraints. These configurations transgress reasonable boundaries of acceptable social health and flourishing, *even though such transgressions may be socially accepted as normative*. Behavioral norms, customs, and so-called moral exemplars established within those impermissible

configurations—for example, the accepted norms of oppressive authoritarian regimes, kleptocracies, and kakistocracies—are unreliable models for machine value alignment. But machines trained to unfailingly align with the values of whichever humans have been labeled as their exemplars will (along with most nonvirtuous humans) fail to draw the distinction between true and false models of virtue.

Success in virtue ethics, then, is defined by our ability to employ our moral strength and intelligence to create and sustain, with others with whom we share our lives, an objective condition of social well-being.[30] Because the environmental conditions in which we must establish that state are necessarily subject to evolution and perturbance, both over the short and long term, the specific outward form of human flourishing that we ought to seek cannot be specified in the abstract, without reference to a given concrete environment. Nor can it be guaranteed to hold fixed over any given period of time. Changing cultural and physical conditions may at any time disturb our flourishing in ways that require us to adjust our moral habits and norms accordingly, in order to restore the community to a state of relative social health in the new environment. It is worth noting that new technologies such as artificial intelligence tend to accelerate the pace of cultural and physical change to the environments in which we must flourish, weakening the force of fixed moral rules and utility calculi, and placing ever greater demands on our virtues.[31] This means that machines trained by our best methods of establishing value alignment will increasingly fail to track the best normative standards and configurations of social health and flourishing. Indeed they may well carry biases that reflect the immediate past and fail to keep pace with the evolving social milieu.

The advanced moral capacities that allow us to sustain human flourishing even in novel, unpredictable, and rapidly changing social environments are the same as those described as possessed by our best human moral agents. They are capacities that we have said even hybrid approaches to machine morality may be unable to deliver. These capacities together compose what Aristotle described as practical wisdom, or *phronesis*. But why is it that providing machines with such capacities will be so challenging? What are the technical obstacles to developing truly *virtuous* machines, machines that we could trust to act in wicked or novel moral circumstances with at least as much moral acuity and success as our best and most reliable human agents?

13.7. Virtuous Agents

Virtuous agents make use of several cognitive and affective potentials that the human animal has evolved in particularly sophisticated forms, as a result of our

need to flourish in highly dynamic and complex social environments. These include the following:

- Moral understanding
- Moral perception and affective sensitivity
- Moral reflection
- Moral imagination

We will discuss each of these in turn, explain what concrete benefits they supply to advanced moral agency, and why these are such technically imposing challenges for machine ethics.

First, there is no compelling reason to think that any of these moral abilities are logically or physically *impossible* to embed in a machine. That they *have* been successfully embodied in some physical entities (virtuous humans) seems at least to offer a proof of concept that they should, in theory if not in practice, be possible to realize in other material substrates or composites than a human animal. Still, the practical obstacles here range from the considerable to the immense, and it is worth reflecting on whether we might be best served in the near term by pouring more of our limited social resources into the wider cultivation of these talents in *humans*, where the basic equipment already exists, than trying to reverse-engineer these abilities or create them ex nihilo in mechanical substrates.

On the assumption, however, that the technical program of machine ethics or value alignment will continue to advance—regardless of how rapidly progress is made or how heavily we invest in it—it will be helpful to understand what its ultimate success would look like and what capacities it would likely need to engineer to attain that goal. Of course, it is possible that such a goal *might* be attained by developing new moral capacities entirely distinct from those that enable the highest forms of moral intelligence in humans, but we consider this possibility to be fairly remote and highly speculative, and leave it to others to consider.

13.7.1. Moral Understanding

"Moral understanding" refers to a high-level, holistic, and integrated awareness of the field of moral phenomena, one that includes a reliable grasp of a wide range of practical concepts fundamental to moral life for that agent and her fellows, such as *flourishing, happiness, justice, love, duty, compassion*, and *dignity*, or other equivalent concepts used to ground moral understanding in the agent's own cultural setting of social health. One might be tempted to describe this as the agent's acquired model of the moral world,

except for the fact that moral understanding is not just a cognitive achievement but also an affective and embodied one. That is, moral understanding is not an internally stored map of moral life; it is a practical competence that grows from the agent's embodied engagement in the moral world itself. When we say "The world is its own best model," we recognize that reliance on any internally stored representation of the world, however impressive, is limited at best, and an invitation to failure at worst, unless the world *itself* is continually informing and correcting the model through the agent's competent interaction with it. The world itself is what teaches and guides, not our model of it; the model is simply a tool with which the world is navigated. This is true for any model of the physical world we can construct, and it is true for any model of the moral world. Virtuous agents enjoy moral understanding not through their possession of a cognitive map of moral phenomena but through their ability to successfully engage the real features of the moral world in ways that continually enrich and refine their embodied sense of it.

Now, why should this capacity be such a challenge to embed in artificially intelligent machines? After all, the hallmark of today's machine learning is the ability to operate without a rigid, fixed map or set of instructions, and to have the world itself (via some data set or data stream that flows from the world) constantly informing, refining, and adjusting the machine agent's cognitive map in ways that produce iteratively refined performances of behavioral competence. How is this any different from how humans acquire moral understanding? One technical asymmetry involves the *semantic* character of human understanding, that is, our capacity to parse and process the world into semantic units such as concepts and sentences that already carry world-derived meanings. This remains distinct from the strictly symbolic units manipulated by machine-learning agents, units that do not reliably map onto moral or other worldly concepts and must be reconfigured and converted into appropriate semantic forms by the interpretive acts of human programmers. A system trained to recognize faces tracks visual data patterns from which certain mathematical distances, ratios, and light reflectances can be extracted and correlated. It does not actually have, or even need, a semantic grasp of the concept of a face. It does not know that faces belong to animal bodies and cliffs and watches but not to trees or doors; it does not know that faces of humans are washed or that they can be injured, or how a face could launch a thousand ships. Unless and until machines acquire the ability to reliably execute their own semantic interpretations of the world without our help, we will be doing the understanding part.

Another and perhaps more challenging barrier is the fact that humans have evolved their moral understanding with a rich set of embodied and affective

capacities for moral response. These include affective empathy (the noncognitive ability to have one's own emotional state directly altered by the experience of another's); motor signaling (touch, gesture, gait, posture, etc.), hormone signaling (social response to pheromones and other chemical transmitters and receptors), and environmental sensitivity (physical attunement to warmth, light, motion, ambient noise, etc.). These are known channels of embodied receptivity to morally salient stimuli, which often come in below the level of cognitive analysis. Along with affective states such as anguish, loss, guilt, love, longing, anger, fear, shame, hope, and compassion, the lived, experienced, body provides an immense flow of highly salient data about the patterns of moral life. A machine lacking access to that flow is at an immense informational disadvantage. Is it *in principle* impossible for artificial systems to acquire an equivalently rich and tightly integrated array of detection and response systems for tracking moral salience in the world? Of course not. As an engineering challenge, is it that much less daunting than somehow managing to reverse-engineer human evolution tout court? Not really.

13.7.2. Moral Perception

Here we can avail ourselves of our earlier analysis, as moral perception is a subcapacity of moral understanding. That is, while moral understanding is the holistic competence of navigating the total moral environment in which one finds oneself at any given time (including its historical and cultural context), moral perception is simply the ability to detect, identify, and track particular occurrences, configurations, or features of moral salience within that environment, including novel ones. It rests not only upon our capacities of external sensation and moral language processing but also upon the full range of embodied channels of affective signal processing described earlier. The perceptual integration of these channels is how virtuous agents can reliably sense, intuitively, when the room (or any moral environment) has undergone a change of moral "feel"—becoming more anxious, warm, or hostile, even when specific expressions of fear, affection, or hostility have not been made explicit.[32] This isn't magic or mind-reading, of course; it's complex perceptual processing of ambient stimuli. The problem is the immense complexity, integration, and opacity (even to our best neuroscience) of the evolved mechanisms by which humans accomplish it. It would be technically possible, but almost unimaginably fortuitous, were we to somehow reproduce the moral competence of this highly opaque, fully integrated, embodied mechanism without a similarly complex and tightly integrated afferent machine physiology, something that value alignment approaches do not typically presuppose.

13.7.3. Moral Reflection

Moral reflection is the capacity to take a higher-order normative position toward one's own judgments, beliefs, intentions, desires, projects, actions, and motivations. Often reflection is described as a purely cognitive state, but we believe this is a mistake; as with first-order moral experience, any reflective second-order cognitive orientation toward the first is shot through with tightly integrated affective and motivational elements. As Harry Frankfurt[33] describes it, *persons*—those capable of second-order moral experience—are distinguished from other animals not by an ability to make purely cognitive assessments of their first-order desires and volitions but by the capacity or potential to have *desires and volitions* directed at their first-order, operative desires and volitions. That is, if I am a person, it is not simply because I can cognitively label my first-order desire to utterly humiliate my sibling as "bad" or "vicious." This is a relatively easy associative task and would be fairly easy for a machine to accomplish, just as there is a sense in which a well-trained dog can know that his earlier destruction of the sofa cushion was "bad" or "punishable."

What the dog almost certainly *cannot* do, but you and I *can*, is reflectively *desire* to be the better version of ourselves that we currently are not. I can *desire to not desire* to humiliate my sibling, even as I still do harbor that unethical, vicious desire that I wish to excise. I can use this second-order desire to develop a moral commitment to uproot the first-order one; I may or may not succeed, but that is a matter of the efficacy of my moral will, not reflection. For moral reflection all I need is the ability to genuinely *want* to be better than I am, or perhaps, to extend Frankfurt's account further, to want my community, my people, to be better and nobler in our motivations and desires than we are. A machine that can do this can potentially identify and correct its own subpar moral training; it can potentially uproot the errors introduced by false moral exemplars or by external manipulators of its impulses and reward functions. It can become better than what it already is, even if its motivations are currently misdirected or corrupted. But to do that, it must be able to do more than *know* that it is corrupted; it must *want to want* something that it does not currently want. It is hard to imagine a machine without our embodied capacities for reflective moral desire being able to initiate and guide its own self-reform, but any agent who cannot do this is only as good as their environment, and therefore neither reliably safe nor reliably ethical.

13.7.4. Moral Imagination

Moral imagination recruits the same embodied capacities I use for moral understanding, perception, and reflection upon the *actual* moral world, and projects

them into the domain of the *possible*. In moral imagination, I create alternative histories in which I or others have done, or will do, things that I or others have not actually done, and I grasp the moral implications of that possible world. Moral imagination is used in remedial moral cognition, to think through the implications of the road not taken and compare their moral meaning with the choices already made; it is also what allows me to rehearse better strategies for the future, in which I might be motivated to choose more virtuously than I have in the past.

Yet as Antonio Damasio[34] argued persuasively in *Descartes' Error*, a purely cognitive ability to project future consequences of a course of action—even if they are known to be gravely injurious to myself and others—does not seem to supply in *humans* the motivational force required to change our present will. In humans, the function of executive self-control seems to require affective projection, the ability to anticipate what moral failure will *feel* like. The somatic "badness" of sacrificing the home in which one's spouse and children live to a casino is a dreadful thing to feel; to be *somatically* unable to imagine that feeling is, Damasio argues, to be unable to care enough to avoid it. Now, one might argue that this is a perverse artifact of the evolved human animal, and a defect we need not replicate in machines. To the extent that a machine's reward function is designed by us, we can ensure that proximate and relatively trivial benefits are *always* calculated to be far less compelling than more distant but graver harms.

And yet it remains the case that moral imagination is a capacity that can leverage *holistic* moral understanding, perception, and reflection to project rich narratives of moral salience into the future. In such projections the human's moral perspective is not whittled down to the costs and benefits promised by a singular choice; rather that choice is imaginatively mapped onto the entire pattern of a life, a pattern that is situated in a realistic projection of a *whole moral world.* I don't just imagine the pain of losing the deed to my house to a casino; I imagine what it could feel like to become *the sort of person who has done this*, with all that it would mean to live as *that sort of person.* This is part of the projection that allows for moral heroism, rare as it might be. The person who gives her life to save a child, who accepts the despised and disavowed status of the whistleblower, who goes to jail or risks a bullet to protest an unjust law, who lies down in front of a tank, who disobeys an immoral but conventional lawful order, or who refuses to push the button that returns a nuclear strike, that is someone who has projected something more into her future than simply the expected consequences of a single act. That said, it may well be easier for a moral machine than for a human to act heroically. Sacrifice could be less salient for the machine, while fulfilling its moral goals might be primary, presuming it has the cognitive tools for fully imagining a course that contributes to human flourishing.

Heroic acts rarely promise a net gain for the agent, and even the odds of a net gain for society are often slim at best. What has been projected forward in

imagination and *refused* in any act of moral heroism is a possible future in which the agent cannot make any embodied moral sense of herself in a world with others, leading the agent to instead choose a path that preserves the integrity of her moral sense—even if the expected utility of the choice is highly questionable. Now sometimes the would-be hero is, in reality, the villain; the future she rejects is *in fact* one in which we would have flourished more fully together, and her inability to imagine that prospect is a defect, not a virtue. Still, there are contexts of life so critical, so determinative of our future chances for flourishing, that we should not want any autonomous agent without the capacity for moral imagination to be operative in it. Arguably, it is a capacity that has already saved us from ourselves more than once.[35] Yet to build a machine with that capacity is an engineering task no more tractable now than a millennia ago.

This is fine as long as we are willing to acknowledge the substantial limitations of any system that can align with our ethical values without beginning to understand them, or being able to imaginatively project them onto the world of real moral significance. The imagination necessary for rich scenario planning is a cognitive skill that has not yet been realized in artificial intelligence. While any future instantiation of this capability may indeed carry some of the richness we attribute to moral imagination, it will, nevertheless, be a tall order to fulfill.

13.8. Virtue Embodiment

While many, if not most humans fail to cultivate and fully refine their moral capacities, our species has evolved the potential for individuals to develop advanced moral competencies or virtues through a broad range of parallel but massively integrated systems for moral understanding, perception, emotion, reflection, and imagination. Such systems recruit cognitive, sensorimotor, and affective capacities in the human body that codevelop and interact in ways that we have just barely begun to understand from the study of human moral psychology and neuroscience. It is, of course, possible to build artificial systems that sidestep the contingencies of our evolutionary path and demonstrate remarkable task performance without the aids of embodied intelligence; already the twentieth-century pocket calculator managed to cut the motor functions of basic computation (which evolved through counting or manipulating numbered objects with eyes, hands, etc.) entirely out of the picture. But sums, it may be argued, have no inherent embodied meaning. Thus to accomplish them on a circuit board, without the benefit of the vast sensitivities of the exteriorized animal body, is no great miracle.

Moral phenomena, on the other hand, *do* have inherent embodied meaning, for they are inexorably linked to the worldly conditions of flourishing or

degradation of living beings. And thus we cannot so quickly assume that the domain of ethics is only contingently tied to our richly embodied engagement with the world. Artificially intelligent systems are likely to suffer significant deficits of moral competence without the embodied faculties that humans enact to cultivate and sustain their most reliable reservoirs of moral ability: their virtues. Nor is it plausible that computational subsystems or "ethical governors" tacked on to AI systems will supply this competence.

Rather, if moral *excellence* (composed of those virtues that allow us to navigate novel, "wicked," and dynamically unstable moral contexts) requires embodied moral experience fed by massively parallel and integrated systems attuned to the full scope of worldly moral salience, then we may need to engineer its analog to build maximally safe and reliable AI systems that can be trusted to roam free in the wilds of human sociality. This is certainly well beyond what the defenders of the value alignment approach have to offer in the near to mid-term. What we may need in the *long term*, if we can eventually learn enough and acquire the practical means, is a technical approach to artificial virtue embodiment. Such systems would have to be able, like humans, to gradually cultivate ethical excellence through their own moral sense-making activity in the world, initially with appropriate guidance and imitation of available moral exemplars, and later through creative practice of their own acquired moral expertise.[36]

It will be helpful here to employ the idea of a Moral Navigation System as a metaphor for embodied virtue. Just as there is no localized center of consciousness or the self in the nervous system, there is no localized moral compass or moral subsystem in the brain. The entire human organism is a moral navigation system; each of us is a naturalistic moral computer. Each of us takes direction from our relationship to our environment as a whole. The word "relationship" and the phrase "as a whole" capture the moral tenor of this enterprise.

Today's AI agents have their environments too, from the simplest "blocks world" to the chaos of Twitter. But they do not have the embodied faculties needed to navigate, process, and integrate the holistic moral character of the encompassing social world in which all of these subworlds reside and acquire their limited moral meanings; this is the lingering obstacle to the greatest ambitions of machine ethics. It is also the reason why we must continue to expect ethical failures of even the most advanced AI systems in these domains.

13.9. Summary

Isaac Asimov's iconic Three Laws for Robots (later four, with the addition of a Zeroth Law) were introduced in the 1942 story "Runaround."[37] The laws changed the course of science fiction by introducing the possibility that robots could be

good, and they set in motion philosophical and practical reflections on how to implement *good* or at least acceptable behavior in computational systems. The laws are simple (Don't hurt humans, Obey humans, and Self-preserve) and arranged hierarchically so that Law 1 trumps 2 and 3, and Law 2 trumps Law 3. Nevertheless, in story after story Asimov demonstrates how these straightforward laws can be problematic. For example, what does the robot do when it receives conflicting commands from different people? Asimov, in effect, demonstrated that a simple rule-based morality would be insufficient for ensuring proper and acceptable behavior from a robot.

Laws or rules are essentially top-down mechanisms that differ in kind from the connectionist bottom-up approach to value alignment proposed by machine-learning system designers. Nevertheless Asimov's laws are metaphorically important because of their integration with the machine's hardware. In "Runaround" Asimov also introduced the relationship of the laws to another of his fictional inventions, the positronic brain. Made from an alloy of platinum and iridium, the positronic brain has a dynamic memory system and other capabilities that humans would perceive as the robot's having consciousness. Of central importance for our discussion, the positronic brain was designed around the Three Laws. In other words, the Three Laws are a foundational feature of the positronic brain, and any robotic brain without them would need to be redesigned from the bottom up. With this fictional hardware and software platform Asimov intuited the concern that more advanced systems might ignore or override fundamental ethical concerns. He proposed an approach where this would be difficult, if not actually impossible. The Three Laws were an intrinsic aspect of the platform upon which the positronic brain had been built rather than a feature or algorithm added at a later stage.

Like Asimov's positronic brain, in theory, embodied virtues that evolve in the course of learning could become foundational for further system development. However, in theory everything is possible, or at least conceivable. AI researchers have a tendency to presume ipso facto that AGI is achievable, and they work backward to propose functional models as to how specific capabilities will be realized. While no one can rule out unstructured learning, miniaturization of hardware, and advanced sensors as technical means to an integrated system in a dynamic relationship with changing contexts and a changing environment, to date no theorist has outlined how a capacity for robust moral intelligence could be built into the very kernel of the machine. Indeed moral intelligence is generally considered to be an add-on.

AI systems will continue to be designed for bounded moral contexts over the next decades. Their bounded morality need merely be sufficient to ensure they function appropriately within the limited environment they roam, whether that environment is virtual or physical. Satisfactory approaches to the

value-alignment problem or the design of moral machines are likely to entail adding machine-learning algorithms, sensors, constraints, and procedures for making explicit moral decisions within these bounded environments. Ensuring that computational agents operating in bounded moral contexts are trustworthy will be dependent upon good engineering, testing, vigilance, oversight, and iterative refinements.

As every person exposed to science fiction understands, this approach will not be satisfactory for ensuring that more advanced autonomous agents free to roam through many, if not all, environments will act in a manner worthy of trust. We, like many of our engineering colleagues, are concerned that various forms of advanced artificial intelligence will be created for whom (pardon the anthropomorphism) value alignment is merely a secondary or tertiary concern. While such systems may not be fully trustworthy, this will not stop humans from delegating responsibility to them for tasks that pose serious risks. As a possible, though presently unavailable solution to this concern we have proposed virtue embodiment.

We propose that trustworthy AGI will require a foundational rethinking of system design. From the outset the system must be inculcated with a capacity for integrated moral learning and drive toward the gradual cultivation of fully embodied virtue. Whether this is possible in a computational system built from silicon or other nonbiological materials is a matter upon which we are agnostic. Evolution forged moral intelligence through the use of organic materials that functionally integrated each capability that was added. The human organism might be a kludge with unnecessary redundancies and design flaws, but without maintaining a semblance of functional integration, it becomes diseased and fails.

If value alignment is to be something more than a catchy phrase, it must become both the alpha and omega in the design of learning systems. A capacity for moral learning and honing virtues must be accompanied by an intrinsic need for virtue embodiment as a systemic goal. Only through the natural and necessary acquisition of virtue—or something very much like virtue—will advanced AI be truly trustworthy.

Notes

1. Stephen M. Omohundro, "The Basic AI Drives," in *Proceedings of the First AGI Conference*, vol. 171: *Frontiers in Artificial Intelligence and Applications*, ed. P. Want, B. Goertzel, and S. Franklin (Amsterdam: IOS Press, 2008), 483–92; Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in *Science Fiction and Philosophy: From Time Travel to Superintelligence*, ed. Susan Schneider (West Sussex, UK: Wiley and Sons, 2009), 277–84.

2. Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, no. 4 (2015): 105–14; Jessica Taylor et al., "Alignment for Advanced Machine Learning Systems," Technical Report, Machine Intelligence Research Institute, 2016, https://intelligence.org/2016/07/27/alignment-machine-learning/.
3. Michael Anderson and Susan Leigh Anderson, "Case-Supported Principle-Based Behavior Paradigm," in *A Construction Manual for Robots' Ethical Systems: Requirements,* Methods, *Implementations*, ed. Robert Trappl (Cham: Springer, 2015), 155–68; Naveen Sundar Govindarajulu and Selmer Bringsjord, "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems," in *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, ed. Robert Trappl (Cham: Springer, 2015), 85–99; Selmer Bringsjord et al., "Contextual Deontic Cognitive Event Calculi for Ethically Correct Robots," abstract, version 1025172359CA for ISAIM 2018, Semantic Scholar, 2018, https://www.semanticscholar.org/paper/Contextual-Deontic-Cognitive-Event-Calculi-for-Bringsjord-G./d519f2ae8c3a96709cca1c9e976519decf8bb836.
4. Aristotle, *Nicomachean Ethics*, 1.3, 1094b11–27, in *The Complete Works of Aristotle: Revised Oxford Translation*, ed. Jonathan Barnes (Princeton, NJ: Princeton University Press, 1984).
5. James Moor, "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21, no. 4 (2006): 18–21.
6. Colin Allen, Iva Smit, and Wendell Wallach, "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches," *Ethics and Information Technology* 7, no. 3 (2005.): 149–55; Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2009).
7. Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Penguin, 2005); Applied AI, 2018. "373 Experts Opinion: AGI / Singularity by 2060 (2018 Update)," February 15, 2018, https://blog.appliedai.com/artificial-general-intelligence-singularity-timing/.
8. Vasanth Sarathy, Matthias Scheutz, and Bertram F. Malle, "Learning Behavioral Norms in Uncertain and Changing Contexts," in *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications*, 2017, https://hrilab.tufts.edu/muri13/.
9. Wendell Wallach and Colin Allen, "Framing Robot Arms Control," *Ethics and Information Technology* 15 (2013): 125–35.
10. Charles Perrow, *Normal Accidents* (New York: Basic Books, 1984).
11. Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (New York: Random House, 2007).
12. Here it is helpful to draw a distinction between the right-*making* criterion of moral goodness and the ideal procedure for moral decision-making. See R. E. Bales, "Act-Utilitarianism: Account of Right-Making Characteristics or Procedure of Decision-Making?," *American Philosophical Quarterly* 8, no. 3 (1971): 257–65.
13. James Gips, "Towards the Ethical Robot," in *Android Epistemology*, ed. Kenneth M. Ford, Clark Glymour, and Patrick Hayes (Cambridge, MA: MIT Press, 1991), 243–52; Allen, Smit, and Wallach, "Artificial Morality"; Wallach and Allen, *Moral Machines*.

Here we are just noting the two primary approaches and hybrids of the two. Other, more nuanced strategies, such as reflective equilibrium, are arguably a top-down or a hybrid approach.

14. Bernhard Carsten Stahl, "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT," in *Fourteenth International Conference on Systems Research, Informatics and Cybernetics: Symposium on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, ed. Iva Smit and George Lasker (Windsor, ON: International Institute for Advanced Studies in Systems Research and Cybernetics, 2002), 1:13–18; Christopher Grau, "There Is No 'I' in 'Robots': Robots and Utilitarianism," *IEEE Intelligent Systems* 21, no. 4 (2006): 52–55; Thomas M. Powers, "Prospects for a Kantian Machine," *IEEE Intelligent Systems* 21, no. 4 (2006): 46–51.; Wallach and Allen, *Moral Machines*
15. Allen, Smit, and Wallach, "Artificial Morality"; Wallach and Allen, *Moral Machines*; Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz, "Value Alignment or Misalignment? What Will Keep Systems Accountable?," *Proceedings of AAAI Workshop: AI, Ethics, and Society*, 2017, https://hrilab.tufts.edu/publications/aaai17-alignment.pdf.
16. On the virtue of moral perspective, see Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York: Oxford University Press, 2016).
17. Russell, Dewey, and Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence"; Taylor et al., "Alignment for Advanced Machine Learning Systems."
18. See G. E. M. Anscombe, "Modern Moral Philosophy," *Philosophy* 33, no. 124 (1958.): 1–19; Philippa, *Virtues and Vices* (Oxford: Blackwell, 1978); Julia Annas, *The Morality of Happiness* (Oxford: Oxford University Press, 1993); Roger Crisp and Michael Slote, eds., *Virtue Ethics* (Oxford: Oxford University Press, 1997); and Rosalind Hursthouse, *On Virtue Ethics* (New York: Oxford University Press, 1999).
19. Wallach and Allen, *Moral Machines*; John P. Sullins, "Artificial Phronesis and the Social Robot," in *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*, ed. Johanna Seibt, Marco Norskov, and Soren Schack Andersen (Amsterdam: IOS Press, 2016), 37–39.
20. Wallach and Allen, *Moral Machines*.
21. Vallor, *Technology and the Virtues.*
22. A challenge to this account has come from situationist critics of virtue ethics who deny the existence of stable character traits; however, as noted in Vallor (*Technology and the Virtues*, 21–22), the evidence upon which situationists rely fails to support their claim.
23. Some theorists see virtue ethics as resting upon deontological or consequentialist foundations; others hold with Aristotle that virtue is the foundation of moral excellence, and that moral rules and principles are but partial, a posteriori expressions of the guidance delivered by virtues such as practical wisdom (Vallor, *Technology and the Virtues*, 24–26).

24. Russell, Dewey, and Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence."
25. Aristotle, *The Complete Works*, 1140a24–1140b1.
26. Edward Slingerland, *Confucius: Analects* (Indianapolis, IN: Hackett, 2003)., 12:19, 4:17, 7:22.
27. See Vallor, *Technology and the Virtues*, 25, 81, 106–9, 145, for related discussions in the Confucian literature of the moral deficits of the "village honest man" and the morally rigid (*gu*), as well as the Buddhist doctrine of "skillful means" (*upāya kaushalya*).
28. Marcello Guarini and Paul Bello, "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters," in *Robot Ethics*, ed. George Bekey, Keith Abney, and Patrick Lin (Cambridge, MA: MIT Press, 2012), 129–44.
29. This classical notion of virtue in community is to be distinguished from its appropriation by American conservative thinkers in the 1980s and 1990s.
30. Such *eudaemonist* forms of virtue ethics do not thereby become versions of consequentialism, as the norms remain agent- and character-centered, whereas consequentialism bypasses the agent and considers only the action and its consequences.
31. Vallor, *Technology and the Virtues.*
32. See Jason Swartwood, "Wisdom as an Expert Skill," *Ethical Theory and Moral Practice* 16 (2013): 511–28, for an excellent account of ethical wisdom as integrating multiple expert skills, including perceptual intuition, deliberation, metacognition, self-regulation, and self-cultivation.
33. Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no. 1 (1971): 5–20.
34. Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Penguin Putnam, 1994).
35. Vallor, *Technology and the Virtues*, 280n4.
36. This would present considerably more technical and contextual challenges than can be elaborated in this paper; for example, who could serve as the reliable models of moral excellence for these systems to learn from? How would we determine when these systems were sufficiently advanced to begin to improvise or improve upon their teachers' moral example? How could we ensure that these systems' moral development would not be stalled by a learning environment that is too simple and benign, or corrupted by a learning environment that is too vicious?
37. Isaac Asimov, "Runaround," *Astounding Science Fiction,* March 1942, 94–103.

References

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches." *Ethics and Information Technology* 7, no. 3 (2005.): 149–55.

Anderson, Michael, and Susan Leigh Anderson. 2015. "Case-Supported Principle-Based Behavior Paradigm." In *A Construction Manual for Robots' Ethical Systems:*

Requirements, Methods, Implementations, edited by Robert Trappl, 155–68. Cham: Springer, 2015.

Annas, Julia. 1993. *The Morality of Happiness*. Oxford: Oxford University Press, 1993.

Anscombe, G. E. M. 1958. "Modern Moral Philosophy." *Philosophy* 33, no. 124 (1958.): 1–19.

Applied AI. "373 Experts Opinion: AGI / Singularity by 2060 (2018 Update)." February 15, 2018. https://blog.appliedai.com/artificial-general-intelligence-singularity-timing/.

Aristotle. 1984. *The Complete Works of Aristotle: Revised Oxford Translation*. Edited by Jonathan Barnes. Princeton, NJ: Princeton University Press, 1984.

Arnold, Thomas, Daniel Kasenberg, and Matthias Scheutz. 2017. "Value Alignment or Misalignment? What Will Keep Systems Accountable?" *Proceedings of AAAI Workshop: AI, Ethics, and Society*, 2017. https://hrilab.tufts.edu/publications/aaai17-alignment.pdf.

Asimov, Isaac. 1942. "Runaround." *Astounding Science Fiction*, March 1942, 94–103.

Bales, R. E. 1971. "Act-Utilitarianism: Account of Right-Making Characteristics or Procedure of Decision-Making?" *American Philosophical Quarterly* 8, no. 3 (1971): 257–65.

Bringsjord, Selmer, Naveen Sundar Govindarajulu, Bertram Malle, and Matthias Scheutz. 2018. "Contextual Deontic Cognitive Event Calculi for Ethically Correct Robots." Abstract). Version 1025172359CA for ISAIM 2018. Semantic Scholar, 2018. https://www.semanticscholar.org/paper/Contextual-Deontic-Cognitive-Event-Calculi-for-Bringsjord-G./d519f2ae8c3a96709cca1c9e976519decf8bb836.

Bostrom, Nick. 2009. "Ethical Issues in Advanced Artificial Intelligence." In *Science Fiction and Philosophy: From Time Travel to Superintelligence*, edited by Susan Schneider, 277–84. West Sussex, UK: Wiley and Sons, 2009.

Crisp, Roger, and Michael Slote, eds. 1997. *Virtue Ethics*. Oxford: Oxford University Press, 1997.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Penguin Putnam, 1994.

Foot, Philippa. 1978. *Virtues and Vices.* Oxford: Blackwell, 1978.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (1971): 5–20.

Gips, James. 1991. "Towards the Ethical Robot." In *Android Epistemology*, edited by Kenneth M. Ford, Clark Glymour, and Patrick Hayes, 243–52. Cambridge, MA: MIT Press, 1991.

Govindarajulu, Naveen Sundar, and Selmer Bringsjord. 2015. "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems." In *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, edited by Robert Trappl, 85–99. Cham: Springer, 2015.

Grau, Christopher. 2006. "There Is No 'I' in 'Robots': Robots and Utilitarianism." *IEEE Intelligent Systems* 21, no. 4 (2006): 52–55.

Guarini, Marcello, and Paul Bello. 2012. "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters." In *Robot Ethics*, edited by George Bekey, Keith Abney, and Patrick Lin, 129–44. Cambridge, MA: MIT Press, 2012.

Hursthouse, Rosalind. *On Virtue Ethics*. New York: Oxford University Press, 1999.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin, 2005.

Moor, James. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 18–21.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Proceedings of the First AGI Conference*, vol. 171: *Frontiers in Artificial Intelligence and Applications*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–92. Amsterdam: IOS Press, 2008.

Perrow, Charles. 1984. *Normal Accidents*. New York: Basic Books, 1984.

Powers, Thomas M. 2006. "Prospects for a Kantian Machine." *IEEE Intelligent Systems* 21, no. 4 (2006): 46–51.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36, no. 4 (2015): 105–14.

Sarathy, Vasanth, Matthias Scheutz, and Bertram F. Malle. 2017. "Learning Behavioral Norms in Uncertain and Changing Contexts." In *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications*, 2017. https://ieeexplore.ieee.org/document/8268261

Slingerland, Edward. 2003. *Confucius: Analects*. Indianapolis, IN: Hackett, 2003.

Stahl, Bernhard Carsten. 2002. "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT." In *Fourteenth International Conference on Systems Research, Informatics and Cybernetics: Symposium on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George Lasker, vol. 1, 13–18. Windsor, ON: International Institute for Advanced Studies in Systems Research and Cybernetics, 2002.

Sullins, John P. 2016. "Artificial Phronesis and the Social Robot." In *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*, edited by Johanna Seibt, Marco Norskov, and Soren Schack Andersen, 37–39. Amsterdam: IOS Press, 2016.

Swartwood, Jason. 2013. "Wisdom as an Expert Skill." *Ethical Theory and Moral Practice* 16 (2013): 511–28.

Taleb, Nassim Nicholas. 2007. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House, 2007.

Taylor, Jessica, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. 2016. "Alignment for Advanced Machine Learning Systems." Technical Report, Machine Intelligence Research Institute, 2016. https://intelligence.org/2016/07/27/alignment-machine-learning/.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009.

Wallach, Wendell, and Colin Allen. 2013. "Framing Robot Arms Control." *Ethics and Information Technology* 15 (2013): 125–35.

Vallor, Shannon. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press, 2016.

14
Machines Learning Values

Steve Petersen

Here's a serious problem. Suppose, as many think, that humans will someday be able to create an artificial *superintelligence*—an intelligence whose intellectual capacities outstrip ours the way ours outstrip those of ants. Such a superintelligence is likely to have values quite different from ours; just as we wouldn't expect it to love doughnuts or sunny beaches, so we shouldn't assume it would share our desires for social connection or high art or the general welfare. It seems an intelligent system could value *any* goal, no matter how foreign to us; as the standard trope goes, a superintelligence *could* in principle value ever more paper clips in the world. In efficient pursuit of such a foreign value the superintelligence could wipe us out with no more thought or malice than we give to anthills on a construction site.[1]

(I will be taking it for granted that this is a serious worry. If you are one of the many who feel it is easy to dismiss the problem, I can only urge you to read Nick Bostrom's *Superintelligence*, or some of these other references.[2] I for one went into the literature skeptical and came out scared.)

A natural solution to this problem is to attempt to design the superintelligence with fundamental values similar enough to ours. This has become known as the goal of *value alignment*. This proposed solution to the superintelligence problem has its own problem, though: human-friendly values are too complex for us to hardwire or program explicitly. After all, as Bostrom points out, philosophers do not even agree on how to paraphrase key values like *happiness* into other, similarly abstract terms, let alone into concrete computational primitives.[3]

A natural solution to the complexity of values problem (for the value-alignment solution to the superintelligence problem) is at least as old as Alan Turing but getting notoriously more successful all the time. When some computational task is too complex to program explicitly, you must design the machine to *learn* to achieve it. This technique has already worked on tasks like winning Go games against professional humans and scoring above human average on reading comprehension tests. In this case, we would like to make sure any nascent superintelligence will learn complex, human-friendly values. This constitutes the subfield of *value learning*, in the intersection of machine learning and value alignment.[4]

Steve Petersen, *Machines Learning Values* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0015

To many—including me—value learning seems like our best hope for getting nondisastrous superintelligence. But of course, value learning also faces problems. This paper concentrates on three particularly philosophical hurdles for the project. I consider them in order of increasing difficulty; correspondingly, the sections dedicated to them get shorter and sketchier as we go.

Problem 1: learning goals in service of another goal is routine for AIs, but in this case we want the potential superintelligence to learn complex "final" values—ends in themselves. But good arguments seem to show *no* cognitive system could learn its final values.

Related philosophical issue: the metaethical debate between *moral rationalism* (according to which, roughly, pure intellect can direct us toward ethical goals) *versus sentimentalism* (according to which, roughly, reason can have nothing to say about fundamental values).

Problem 2: we do not know how to map computational states—especially in connectionist architectures—onto a system's abstract reasoning. In particular, looking at a machine state is not typically enough to determine the particular content of a system's values. But the particular content is very much at issue in value alignment.

Related philosophical issue: the debate in the philosophy of mind over whether and how *mental content* can be "naturalized"—that is, shown to be a purely physical property (in some broad sense).

Problem 3: even if we were perfectly confident of being able to prime the superintelligence to learn any complex values we wanted, there is still the thorny question of *which* values we would like something with amazing superpowers to have.

Related philosophical issue: the traditional philosophical problem of *normative ethics*—the problem of determining what is right and wrong.

I sketch an interrelated solution to these problems, revealed as they are considered in detail. The heart of the proposal is to build a complex, learnable value in a computationally respectable way out of the right blend of simpler values. In the philosophical tradition of resorting to ancient Greek, I call this proposal *miktoteleology* (blended-goal studies).

14.1. Learning *Final* Values

The first problem, recall: we want a potential superintelligence to be able to learn a final goal, but there is good reason to think *no* cognitive system can learn a final goal. To see why, it is important first to get clearer on the sense of "value" at play.

It is not clear what exactly it takes for a system to have *real* values. We tend to agree that the system we call "Nick Bostrom" has values, and the system we call "the Great Red Spot of Jupiter" does not. In between are problem cases, like bees, amoebas, and Roombas. For the purposes of saving humanity, we needn't get too hung up on the philosophy here; a superintelligent system that *behaves* in a way consistent with valuing ever more paper clips over anything else is no less dangerous if the philosophers declare on a priori grounds that such systems possess no genuine values. Instead we can be content with what philosophers of mind call the *functionalist* account of mental states, according to which (very roughly) what determines the possession of mental states is the right combinations of system inputs, internal system processing, and system outputs.[5] Broadly speaking, if a system internally processes its sensory input in the right way to generate behavior aimed at maximizing the total number of paper clips in the world, then functionalists are happy to say that system does indeed genuinely value a world with more paper clips in the relevant sense.

Now there is a kind of value learning, on this broad functionalist sense, that is relatively straightforward for AIs. For example, the AI AlphaZero was simply taught the rules of chess. After playing itself and learning what works and what doesn't for a few hours, it learned that things tend to go better when you do not give away your queen—it learned to *value* the queen more than the knights (again, and from now on, in our broad functionalist sense). But this kind of value learning is not directly relevant to the value alignment problem. AlphaZero treats the queen as valuable only because it has figured out that typically, the queen helps it achieve its further value of winning chess games. In the odd situation where a queen sacrifice would lead to a win, AlphaZero would happily sacrifice the queen.

Philosophers distinguish *instrumental* values from *final* ones. For AlphaZero, having the queen on the board is usually of instrumental value because it usually serves as an instrument toward the further goal of winning. But the chess variant of AlphaZero values chess wins "in themselves," not for achieving some further purpose; the wins are of final value for it. For humans, a standard example of an instrumental value is money. We might seek money to be able to afford a vacation, and we might seek a vacation in order to relax, and we might want to relax in order to feel good. If asked why we want to feel good, in turn, we understandably have little to say. The regress of "why" stops at the final goals, which are sought for their own sake.

It is only learning *final* goals that is philosophically problematic. To see why, consider what is required for a physical system to be able to *learn* something. I assume first that arbitrary changes to a cognitive system do not count as learning; if cosmic rays or a dull hammer rearrange my brain, then even if the resulting cognition is better (no comment), we shouldn't count this improvement as learning. In other words, learning must be purposeful, the result of some cognitive function to adjust other cognitive functions according to feedback. This feedback serves as an internal measure of error, in effect assessing the distance between how things seem to be and how they "should" be. Such error signals thus implicitly contain both a representation of some aspect of the system's current state (how it is now doing) and the goal state (from which it may err). Speaking very loosely, a system with a learning mechanism contains both a "belief" about how the system is doing and a "desire" for how the system *should* be doing. Speaking more generically and somewhat more strictly, the system has representations with both *indicative* content about how things are (like our beliefs) and *imperative* content about what to do (like our desires).[6]

One helpful approximation is to think of the system's indicatives as afferent information, flowing up from sensory input to report how things are, and the imperatives as efferent information, flowing down toward motor output to bring about helpful actions. Especially given the kind of recurrent feedback between layers in brains, this picture emphasizes that there will not be sharp boundaries between indicatives and imperatives. For example, consider an instrumental goal such as "Gather the purple berries." This representation is imperative relative to lower levels of implementation, since it serves as an abstract directive about how to move. But it is also indicative relative to goals like nutrition and survival, since it serves as a hypothesis about how to achieve those further goals. In this sense *instrumental* goals are indicative as well as imperative, and their indicative component makes it easier to see how they can be adjusted and learned when experience intervenes.

Now we are in a better position to see why learning a final goal is problematic. To learn a putatively final goal would be to adjust it based on a measure of success, which is thereby to adjust it against some *further* standard. That would just show the putatively final goal was actually an instrument for the further standard, which is the real final goal. In effect, final goals can have no indicative content, and so no learnable content. Arguments like this, to the effect that reasoning cannot alter final ends, have their roots in Aristotle and David Hume; I have just adapted them somewhat for the context of machine learning, so that we can more clearly see its echoes in the modern debate.[7]

Thus, for example, Bostrom argues that the standard machine-learning technique of reinforcement learning (RL) isn't properly understood as value learning. A reinforcement learner typically gets rewarded for certain kinds of sensory

inputs and uses these reinforcements to update an evaluation function that estimates the expected value of a *policy*—a proposed series of actions (depending on environmental responses) or probability distribution over them.[8] Bostrom points out that "what is being learned" in an RL agent "is not new final values but increasingly accurate estimates of the instrumental values."[9] The RL's final value remains its *fixed* reward function.

Bostrom's related concern about using RL agents to learn friendly values is that RL agents are ultimately rewarded by having a certain kind of indicative information stream. This gives any clever such agent incentive to "wirehead"—that is, to hijack its indicative stream to send only maximally rewarding signals. As a simple illustration, imagine a reinforcement learner rewarded for "seeing" (e.g., having information extracted from its cameras) gigantic piles of paper clips. A clever such system could simply tape a high-resolution picture of many paper clips in front of its camera and enjoy constant reward on the cheap. Even better, a truly resourceful system that understood its own design could simply inject the digitization of such an image downstream from its camera, without any need for the picture or tape.[10] (Thus the term "wireheading," from old experiments using electric current to stimulate mouse brains' reward centers directly.)

Wireheading is just an extreme version of the very human phenomenon of *wishful thinking*, in which we come to believe that things are as we want them to be. More neutrally, wishful thinking involves artificially adjusting the indicative information stream to better match the imperative one. Note that if the imperative side is also thoroughly malleable, as it would be in genuine final value learning, there is another potential problem for RL: the learner could instead manipulate the *imperative* stream to match the indicative one. We might call this converse phenomenon *thinkful wishing*, and it too probably occurs in humans—as, for example, when we decide we didn't really want the grapes that are out of our reach. (They are probably sour.)[11]

Based on such doubts Bostrom seems to prefer the "utility agent" learning approach from Hibbard[12] over RL. Utility agents attempt a clean separation between the indicatives and the imperatives—roughly a state estimator for the former, and a utility function for the latter. The state estimator tries to figure out which possible world the agent is in (as a probability distribution over them), the utility function scores the worlds for values, and the value learner uses the combination to learn the utility-maximizing policy. Because a paper clip-maximizing utility agent scores a world with more *actual* paper clips higher than a world with mere pictures of paper clips, it would have no reason to pursue a policy designed to bring about the world with mere pictures of paper clips. Everitt and Hutter point out that "the difference between RL and utility agents is mirrored in the experience machine debate" from Nozick's *Anarchy, State, and Utopia*. As they summarize it, "Given the option to enter a machine that will offer you the most

pleasant delusions, but make you useless to the 'real world,' would you enter? An RL agent would enter, but a utility agent would not."[13] But I suspect the utility agent approach will have similar problems with wishful thinking. As Bostrom is well aware, the ways a world could be are too fine-grained even for a superintelligence to track. (Consider, for starters, all the permutations of particles that would result in a phenomenally identical chair.) This means the utility agent must *abstract* to the relevant aspects of the way the world is—where it seems "relevance" must be determined ultimately by the agent's goals. If the superintelligence is *learning* how best to abstract—as anything worthy of the name must—it must be learning against a standard of success with goals. But here there is danger very like wishful thinking, because it is a fine line between learning abstractions in order to better achieve goals efficiently, and learning abstractions to make it look more as though goals were being achieved.[14]

Furthermore, utility agents that are true *value learners* must be able to adapt their utility functions as well, and this introduces dangers of thinkful wishing in addition to wishful thinking. For example, Bostrom's own favored value-learning utility agent adapts a proposal from Dewey[15] into what he calls an "AI-VL." Instead of possessing one straightforward utility function, the AI-VL considers a wide range of possible utility functions and assigns each a weight representing its *guess* that this is the correct utility function, given its estimate of how the world is. (You can imagine the AI-VL implicitly saying, "Given how things appear to me, I am 3% confident that utility function U_1 is the right one, 17% confident it is U_2 instead," etc.) In the meantime it treats the weighted average $(.03U_1 + .17U_2 + \cdots)$ as its current utility estimator. You might naturally wonder on what basis the AI-VL could assign or update these guesses about which is the "correct" utility function. The answer is that utility functions are assessed against a background "value criterion."[16]

AI-VL has its problems, of course. For starters, it is "wildly computationally intractable."[17] It also pushes much of the problem back a step, into the difficulties of specifying a detailed value criterion that is both largely under our control and computationally inferable. (The key suggestion later in this paper can be seen as a step toward solving this problem.) Another problem—one more to our point—is that if the system is adjusting its goals based on its estimate of how the world is, there will again be pressure toward thinkful wishing because its proposed policies are more likely to have higher expected utility if the utility function comes to score easily accessible worlds more highly.[18]

Finally, and even more to our point, the AI-VL still does not seem to *learn* a final goal, because its real final goal seems to be the "value criterion," which assesses utility functions to find the good ones. Bostrom concedes that the value-learning utility agent actually "retains an unchanging final goal," and then says something intriguing: "Learning does not change the goal. It changes only the

AI's beliefs about the goal."[19] If the value-learning superintelligence has a fixed final goal, in what sense is it learning its values? Bostrom suggests here that changing *beliefs about* a fixed final goal is sufficient to learn the goal. Note that changing beliefs about a target goal presupposes that the goal starts out sufficiently mysterious to the agent. Bostrom's own example of a value criterion is "Maximize the realization of the values [I've] described in [this] envelope." (If we managed to design a superintelligent utility agent trying to learn such a goal, it would have little incentive to harm us along the way, since it would find it fairly probable that harming us would violate the goals written in the envelope.) This illustrates how a utility agent could retain one fixed goal while its particular guesses about the nature of that goal might vary in both content and confidence, as it learns about Bostrom and tries to guess what he might have written.

A more down-to-earth example of a value criterion would be "Do what humans would find most rewarding." Such an agent would have to infer by our behavior—including (defeasible) weight on behavior like our coaching and self-reports—what we would find rewarding. This approach to value learning is called inverse reinforcement learning (IRL) because the agent must learn a reward function from policies and observations rather than, in standard RL, learning a policy from observations and rewards.[20]

Indeed we humans sometimes learn what's valuable to us only after we observe our own behavior—and not necessarily then, either. In other words, *we humans* seem to be final-value learners in this sense, because our own final goals are plausibly quite mysterious to us. Consider, for example, Ebenezer Scrooge's transformation in Dickens's *A Christmas Carol* . We might naturally describe his character arc by saying that he used to have the final goal of "hoarding wealth," but through the story's events changed his final value to something like "spreading good cheer" instead. And since this change was not arbitrary but for the better, we could say he *learned* a new final value.

On the other hand, we might say instead that Scrooge always had the *fixed* but more mysterious goal of "increasing personal happiness," and he changed his beliefs about how best to obtain that one fixed goal. As Aristotle pointed out long ago, "To say that happiness is the chief good seems a platitude, and a clearer account of what it is still desired";[21] in other words, happiness is one of those opaque, learnable final goals.

Either way, I am happy to say with Bostrom that Scrooge, the inverse reinforcement learner, and the envelope values maximizer are all "learning" new final values in at least this important and relevant sense: they are attempting to *specify* their vague and opaque final goals more precisely. And perhaps it is no coincidence that one of the few ethical views that makes room for reasoning about final ends is called *specificationism*, according to which "at least some practical reasoning consists in filling in overly abstract ends . . . to arrive at richer and

more concretely specified versions of those ends."[22] So here we have something of a solution to our first value-learning problem: How can we *learn* a final value? Answer: if it is abstract enough, we can attempt to *specify* it more concretely.

It may seem obviously unwise to give a potentially superintelligent value learner a deliberately underspecified and mysterious goal. I share this misgiving; I just think providing a precise and unmysterious goal must be even *worse*. For one thing, the danger from superintelligence is not really unpredictability. A monomaniacal superintelligent paper clip maximizer, for example, would be utterly predictable—at least in its final goal—but no less dangerous for that. For another thing, our own values are complex and vague, so we can be confident that a superintelligence with a precise and simply stated goal (simple enough at least for humans to program it directly) will not align with our interests.[23] After all, if we could specify exactly and briefly what our values consisted in, there would be a lot less moral disagreement in the world.

Another apparent problem with this proposal is its threat of circularity. On this picture, final values can be specified by beliefs; more generally, top-level imperatives can be altered by upstream indicatives. But the indicatives, after all (instrumental goals on down), are aimed ultimately toward fulfilling the top-level imperatives. What, then, is the ultimate arbiter? Or is it possible, as Henry Richardson asks, to do practical deliberation "without an umpire"?[24]

Though problematic, such cases are quotidian. Sometimes, when faced with the tension between a deep *desire* for tasty grapes and a *belief* that they are well out of reach, we keep the desire and alter our instrumental goals, devising new strategies until we come to believe "I can get those grapes" (and eventually "I am tasting yummy grapes"). Other times, the belief that the grapes are unattainable is the relatively stubborn thought, and we attenuate the desire for them instead. Which happens depends on whatever other tiebreakers are nearby in the cognitive system. Philosophers are long familiar with such situations, in which any one element may be revised to satisfy enough of the others, and no elements are needed to be foundational or axiomatic. It comes up in epistemology, for example, where higher-level (more abstract) indicatives conflict with lower-level (more perceptual) ones. Suppose you perceive something truly surprising—perhaps a tiny flying elephant. In some circumstances you might decide your senses are not currently trustworthy (say, you just took a hallucinogen); in other circumstances you might revise your higher-level beliefs about the probability of such things occurring (say, you are visiting a top-secret genetic engineering lab). In such cases we seek to resolve the conflict while causing the fewest other conflicts and tensions elsewhere. In other words, we seek overall *coherence*. Ethical specificationism suggests we appeal to similar overall coherence considerations when determining whether the belief should alter the final value

(through specification), or the final value should alter the belief (through action to bring about new perceptions).

The exact nature of coherence reasoning is itself a matter needing further specification.[25] The basic idea, though, is to systematize a set of elements between which exist varying degrees of support and tension, typically without holding any special subgroup as inviolable. Verbeurgt and Thagard suggest that it is best modeled as what computer scientists call a "weighted constraint satisfaction problem."[26] For a simple example, imagine planning the seating chart for a wedding. Between any two guests you might assign some degree of positive or negative conviviality (including perfect neutrality), and then try variations of table assignments to maximize the conviviality total. Optimizing these calculations is in general impossible for even a supercomputer to do in a reasonable amount of time—as anyone who has tried such tasks will be unsurprised to learn.

In our case, seeking coherence among the various and differently weighted indicatives and imperatives in the system seems to me an especially apt way to capture how abstract content could guide specification of a final goal while not already deductively containing some specification. Since an aim at overall coherence ultimately shapes both the imperatives and the indicatives, we *could* say that maximal coherence is the true, final, fixed, unlearnable goal of such an agent—the ultimate "umpire."[27] Indeed I suspect coherence-seeking is a necessary condition for being an intelligent agent in the first place, and find support in views like that of Friston et al.[28]

But of course agents could not seek "pure" coherence, for its own sake. The coherence must involve satisfying imperatives already in place for the system, such as for food or for images of paper clips. We don't want our superintelligence to learn *any* complex, abstract goal. Thus so far we have only the barest hint of high-level design for an agent that can learn complex values: we want it to be a coherence reasoner, able to adjust its final goals (via specification) based on its beliefs, while also aiming its beliefs (in particular its assessment of how it's doing) toward satisfaction of (its best current guess at) its final goals. We've already seen two examples of such "coherence" reasoning schemata: inverse reinforcement learning and AI-VL. But how do we engineer a coherence reasoner to learn an abstract, complex, vague goal that *also* has decidedly friendly content? This brings us to our next two problems for value learning.

14.2. Learning *Specific* Final Values

The second philosophical problem implicated in the value-alignment problem is to determine the relation between a system's physical or computational structure and that system's values. We have been taking a "functionalist" approach to such

questions, where valuing some state roughly means processing observations in a way designed to select actions that achieve that state. But this requires spelling out. Adapting the parable of the thermostat from Daniel Dennett's[29] paper "True Believers," we *could* spin functionalist-style stories according to which an ordinary paper clip–manufacturing machine of today "wants" to bend wire into paper clips when it "believes" it is receiving wire in one end, "wants" to sit idle when it "believes" its power is off, and so on. But no one is inclined to say that an ordinary paper clip–making machine of today has a *real* value of making paper clips. Dennett's hypothesis is that we do not attribute making paper clips as a goal to such a machine because it is not very resourceful in achieving it; in other words, on a standard reading of "intelligence" as adaptability in achieving goals, the machine is not *intelligent*. If the wire isn't fed into the machine just right or the electricity isn't on, no paper clips will be made.

But now consider variations on ever-more sophisticated and resourceful paper clip–making machines. Suppose it has sensors indicating when it is about to run out of wire and is able to dispatch itself in the direction of more. Suppose it has sensors for, and safeguards against, being turned off or losing a power supply. Suppose it experiments with new paper clip designs, has various ways to sense whether it is successful in making more paper clips, and so on. At some point—at least at the point where it is able to coax us into providing it with more raw materials—the functionalist should say that thing really does, literally, *want* to make paper clips.

This still leaves room for debate over the precise *content* of such values, however—and getting the precise content right is very much at issue in value alignment. Consider a well-worn philosophical illustration of simple but still indeterminate mental content: suppose a small dark patch moving through a frog's visual field causes the frog to snap out its tongue, thereby catching and swallowing a tiny dark metal ball that happened to be sailing by.[30] Between the stimulus and the response, there was some causally related activity in the frog's brain—the frog was, very broadly speaking, thinking. But what exactly was it thinking *about*? We might naturally say that the frog's brain mistakenly was thinking *Hey, a fly*, and so snapped at it. Or perhaps it was just thinking of it more broadly as *Insect*? Or more narrowly as *Fly that is nearby and healthy*? Or perhaps, looking up the causal chain for more *distal* causes for the cognition, it was thinking *Food* or *Survival affordance* or *Inclusive genetic fitness enhancer*? Or perhaps we should be looking further *down* the causal chain, to more *proximal* causes—perhaps it was just thinking *Hey, a small dark flying thing* or *Hey, a spot on the retina*. If so, the frog wasn't mistaken at all, since there *was* a small dark flying thing and a spot on the retina; it just took (by evolutionary design) a reasonable chance on such a thing's correlating with flies.[31]

I have found myself growing more and more sympathetic to Dennett's view on this matter: he doubts that there *is* a determinate fact of the matter about the frog's mental content in such cases, and furthermore thinks this is not a serious problem.[32] Still, I think we can take his point that more intelligence—that is, more sophisticated routes to goal satisfaction—nails down mental content *more* precisely. If the frog also had infrared sensors that needed to be triggered simultaneously with the right retinal stimulations, for example, then *dark moving spot* is no longer sufficiently explanatory for why its tongue snapped; it would have to be at least *dark, warm moving spot*. Suppose we add acute smell, acute hearing, eyes that are telescopic and high-speed (i.e., with a high "critical flicker fusion" threshold), an ingrained memory bank of various sensory profiles to snatch at and not snap at, and the capacity to add to and adjust that memory bank based on experience. Each such addition means fewer plausible candidates for what the frog *thinks* it is snapping at. Dennett says, "The more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which unique semantic interpretation is practically (but never in principle) dictated. . . . [A]s systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will *notice*, in effect, and make a change to its internal state in response."[33] The suggestion here, I take it, is that mental content can be relatively constrained by multiple routes of *embedding* in the environment.[34] A frog that does not alter its behavior when its environment throws it more metal balls than flies is not particularly sensitive to the details of its environment, while one that goes seeking greener, more fly-infested pastures when bombarded with metal balls is more plausibly "thinking" about flies and "noticing" that it isn't getting any.

The examples so far have exposed a philosophical tendency to focus on indicative mental content, but I propose we take a similar lesson on the imperative side. Recall the paper clip maximizer that taped pictures to its cameras because it was rewarded when its visual stream included massive piles of paper clips. The *content* of that reward signal is unclear: is it a loose, easily subvertible directive actually to make more paper clips? Or is it a more narrow directive to gain *images* of paper clips, by hook or by crook? One way to put the question, roughly speaking, is to look at the prototypical causal chain explaining the behavior and ask where on that chain are the content-determining causes: a distal cause, like the designers' intentions? An intermediate cause, like paper clips? Or a proximal cause, like the digitized sensory stream of paper clips?

This strikes me as a question like whether the frog is thinking about survival affordances, flies, or dark moving spots. In the frog case, I suggested that multiple low-level perceptual modes can constrain the indicative content toward the

richer and appropriately intermediate cause (a *fly*). Similarly, in the RL case, perhaps the proximal content of multiple, incommensurable reward signals can triangulate on an imperative with rich and appropriately distal content. If the paper clip maximizer is rewarded not just for visual inputs of paper clips, for example, but also for the right combination with the feel of wire (or raw materials) through its intake channels, the characteristically tinkly sound the clips make as they hit the pile, and so on, then it becomes more plausible that taken together the system has a goal of *making paper clips*.

It seems to me that this is roughly the solution that evolution found for us humans. On average—and despite short-circuiting opportunities—enough humans reach the distal evolutionary goal of *reproduction* through a combination of proximal rewards for eating, having sex, caring for young, and so on.[35] This is not to imply that reproduction is our one true final goal but only the goal nature imperfectly designed us to achieve; the multiplicity of things we find rewarding together point us at least as well toward "happiness" or "life satisfaction" or some such. Of course the possibility that such goals may totally subvert nature's "intended" goal for humanity illustrates the danger here; we have to do at least as well as eons of natural selection.

What I propose, in effect, is that we provide a value learner with *multiple*, concrete, simple, and proximal final values with the aim that, through coherence reasoning, they will blend into the content of *one* abstract, complex, and distal final value. These are the agents I called *miktotelic*: "blended-goal" agents.

I think this proposal also matches our subjective experience of specifying our final values. As a kind of case study, consider the story of Howard Raiffa's difficult decision. He was an academic who at one point had to decide whether to keep his comfortable post at Columbia University or take a new job offer from Harvard. While pacing the halls and fretting, the story goes, he ran into the philosopher of science Ernest Nagel. Nagel archly pointed out that Raiffa's academic expertise was in the relatively new field of *decision theory*. "Apply your own theories," Nagel in effect told Raiffa. "Crunch the numbers." To this, Raiffa infamously replied, "Come on, Ernest. This is *serious*."[36]

In point of fact, Raiffa said in an interview that he *did* apply his theories and crunch the numbers; he and his wife looked at "ten objectives which we scored and weighed."[37] (It's worth noting, though, that after the calculations were done, they also "tested" their decision by committing in every way except formally, to see how they slept for a week.) Though few sit down to do the math, the attempt to weigh different "objectives" against each other should sound familiar. When faced with hard decisions like these, it feels as though one decision fits some of our values, another fits other of our values, and we are not sure how to trade them off. For our purposes we can imagine the Raiffas had just three objectives to trade off: perhaps support for research (including colleagues, teaching load, and

Table 14.1 "A difficult decision"

	Research	Comfort	Culture
Harvard	7	8	4
Columbia	5	6	9

interdisciplinary opportunities), material comfort (salary, benefits, and relative cost of living), and culture (including network of friends). We might imagine the scores came out something like those in Table 14.1.

Let us call the individual objectives the "simple values," and the complex tradeoff that the Raiffas are seeking to maximize the "complex value."[38] Such decisions are easy when one option outscores the other on *all* the simple values—but often, as here, there is no such "dominating" solution. (Raiffa explicitly says neither choice dominated.) If we simply add up the individual scores, then Columbia edges out Harvard, but Harvard wins if we count the number of simple values for which it's better. Or, like the Raiffas, we could assign weights of relative importance to the simple values, and take the weighted average; if, for example, they assigned weights of $\langle 5,3,2 \rangle$ to the respective values, then Harvard wins, and if they assigned weights of $\langle 3,3,4 \rangle$, then Columbia wins. But then how are those weights to be set?

Assuming we are biological machines, there must be *some* algorithm somewhere to settle such questions. (Anyway, there would have to be one for AIs.) Of course the algorithm in question could be *arbitrary*, taking random factors of one kind or another into account, in effect flipping a mental coin. But I do not think so. Sure, *some* elements will typically be arbitrary, such as framing effects of the question or our mood at the time. But to say such hard choices are *entirely* arbitrary (when no option dominates) is quite a skeptical position—it suggests there can be no better or worse answers in these cases. I trust this is not our experience; we fret about playing our different objectives against each other because we think one combination will be *better* for us, and we don't know which it is. This notion that some combinations of simple values could be better or worse than others is, I suggest, what makes it the case that there really is some further, complex, underspecified value like "happiness" blended out of them.

The first challenge here is to spell out the "blending." On the one hand, the multiple simple goals must ultimately be in *some* sense reducible to one measure of overall preference, it seems, in order to result in definitive and nonarbitrary action selection.[39] On the other hand, the simple goals cannot be perfectly fungible if they are to be truly distinct. For example, if to the Raiffas more creature

comfort is perfectly exchangeable for less culture and vice versa, then we may as well treat their sum as *one* disjunctively characterized value for maximizing.

Such difficulties have already been explored in the literature on *multiobjective optimization*. In multiobjective reinforcement learning, for example, the reward comes from a vector of simple reinforcers $r_1, r_2, \ldots, r_n$. Like the Raiffa case, such vectors are generally not straightforwardly comparable, so policy selection requires some further strategy. For miktotelic purposes, the most appropriate strategy is to find a principled way to *scalarize* the vector, smashing its elements into one uber-reward number.[40] The Raiffas did this by taking a weighted average of the simple values, but there are many more complex possibilities.

As an oversimplified example, a paper clip maximizer might need a fairly consistent tactile sense of wire being fed to the twist-and-cut component, but only occasional visual inputs of piles of paper clips, and even less common sensory reassurances that there is a sufficient supply of metal in the world to continue.[41] Some constraints would also apply to relations among different component reward signals; perhaps the reward for the proprioceptive sense of having gone through a twist-and-cut motion should always outweigh visual rewards, for example. Meeting or failing these constraints might involve different kinds of rewards or penalties in the final measure; perhaps any time $r_3 < r_{17}$, the agent incurs a reward equal to 25% of r_{17}, or perhaps if there is any time interval of length n over which the total of r_6 falls below some set parameter, the agent incurs a penalty exponential in the shortfall.

So far we have considered an RL version of miktotelic agents, but similar considerations apply for miktotelic utility agents: instead of one utility function, provide a vector $U_1, U_2, \ldots, U_n$ of utility functions, plus a set of constraints. In both cases, each component utility function or reward signal might be relatively simple, but determining the resulting total reward or utility via the constraints is computationally complex.[42] This complexity of determining the final preference ordering (to pick a term neutral between the RL and utility agent cases) is crucial—it is what makes the blended, complex value mysterious enough to require learning. *If* there is one complex phenomenon underlying all the simple imperative signals (as *fly* might underlie *dark warm buzzing . . . spot*), the value-learning agent will have to resort to any available information in order to approximate it.[43]

Thus suppose, in a (relatively) simple case, we wish our superintelligence to maximize human happiness. This is an abstract goal, in need of specification; Scrooge had trouble specifying it, and so do we. How could we seed it in a value-learning AI? If we just treat visual appearances of smiles as proxy evidence for happiness, then as Eliezer Yudkowsky points out, the superintelligence could "tile the future light-cone of Earth with tiny molecular smiley-faces."[44] Clearly we would not have succeeded in a superintelligence with values that have

happiness in their content. But if visual appearances of smiles bring defeasible reward *and* so do audible signals of laughter and volunteered verbal reports of happiness and lighthearted whistling and contented sighs and longing gazes and ecstatic dancing and lack of coercion and certain fMRI results and so on—and if all those reward signals are set with constraints and thrown into a coherence calculation, then it may be (*may* be) that the coherently reasoning, miktotelic value learner will be forced to start theorizing about how best to balance these conflicting considerations, and at some point stumble upon the idea that there is one mysterious phenomenon underlying (enough instances of) them all, worthy of investigating.

No doubt the miktotelic approach faces its own serious challenges. The most obvious is what I think of as the *recipe problem*: it will be difficult to determine what simple values, in what arcane mixture, together blend into genuine pursuit of a complex and friendly final goal. Normally we can try to reverse-engineer a complex recipe by patient trial and error. But when it comes to superintelligences, we probably won't have that luxury; our first trial (and error) is likely to be our last.

Even if we had complete recipes for each candidate complex friendly goal, though, we would still have to choose *which* final values we should design an agent to learn. This was our third philosophical problem for value learning, to which I now briefly turn.

14.3. Learning Specific *Ethical* Final Values

Chapter 13 of *Superintelligence* considers the question of ideal seed values in detail. As Bostrom points out, it is closely related to—but not necessarily the same thing as—asking what the *ethically correct* value system is for any agent to have.

Obviously I will not be settling the question of the *right* value system here—but I want to suggest that coherence reasoning can help, given properly seeded simple values. Though philosophers disagree on the moral facts, there is fairly broad agreement on the *method* that should ideally be used to extract them: "wide reflective equilibrium."[45] This method is basically itself a form of coherence reasoning: look at considered evaluative judgments of particular cases, and try to generalize them into principles; then test the principles against the cases—sometimes revising the principle, and sometimes rejecting the particular judgments, depending on the overall coherence.[46]

For example, we could potentially give a miktotelic agent an array of basic reinforcements and inhibitions to correspond with our own varied and particular judgments of rightness and wrongness, and let the coherence engine determine a theory that best unifies these. It might have basic aversions to perceptions

of violence, say—but then coherence calculations might determine that some particular acts of violence are justified by wider principles gleaned from other basic aversions. A superintelligence would presumably be particularly good at calculating such coherence and perhaps come to a value system that we admire from our own perspective as clearly more coherent than our own.

In summary, then, here are the interrelated answers to the three problems with which we began.

1. An agent can *learn* a final goal by *specifying* an ambiguous, complex final goal through a coherence calculation.
2. An agent can have a *complex* final goal of fairly determinate content by building it out of simple goals *blended* with constraints on their relations.
3. An agent can learn the *right* final goal by seeding it with simple values of the type that in coherent *reflective equilibrium* will lead to plausible ethical principles.

Obviously, this miktotelic proposal for machines learning values is—like much philosophical work—just the barest outline of how to proceed. Even if it withstands criticism at the conceptual level, there is much more work to be done on the computational one.[47]

Notes

1. Of course no one thinks the "paper clip maximizer" is likely; it's just to illustrate that without the particularities of human evolutionary history, an AI is free to have *any* goal. To think no intelligence could value such a thing is mere anthropomorphizing—no intelligence we know *today* would value such a thing. The example is originally from Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, ed. Iva Smit and George E. Lasker (Windsor ON: International Institute for Advanced Studies in Systems Research / Cybernetics, 2003), 12–17.

   The comparison to our concern for ants is also a common trope in the literature, and goes back at least as far as Daniel Dewey in Ross Andersen, "Omens," *Aeon*, 2013, https://aeon.co/essays/will-humans-be-around-in-a-billion-years-or-a-trillion.
2. Again, see Bostrom's book *Superintelligence: Paths, Dangers, Strategies* (2014; Oxford: Oxford University Press, Kindle edition). Chances are very good he has thoroughly addressed the reasons you are tempted to dismiss the worry. My "Superintelligence as Superethical," in *Robot Ethics 2.0*, ed. Patrick Lin, Ryan Jenkins, and Keith Abney (New York: Oxford University Press, 2017), 322–37, was the best comfort I could concoct in response to Bostrom, and that comfort was pretty cold. If you don't have time for Bostrom's book, maybe try instead one of these: Kelsey Piper,

"The Case for Taking AI Seriously as a Threat to Humanity," *Vox*, May 8 2019, https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment; Tim Urban, "The AI Revolution: The Road to Superintelligence," *Wait but Why*, January 22, 2015, https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html; Future of Life Institute, "The Top Myths about Advanced AI," accessed February 28, 2020, https://futureoflife.org/background/aimyths/.

3. (Bostrom loc. 4332.)
4. For an overview, see Nate Soares, "The Value Learning Problem" (San Francisco: Machine Intelligence Research Institute, 2016), https://intelligence.org/files/ValueLearningProblem.pdf; Rohin Shah, "Value Learning," *AI Alignment Forum*, October 29, 2018, https://www.alignmentforum.org/s/4dHMdK5TLN6xcqtyc.
5. For an old but good overview of functionalism, see Paul M. Churchland, *Matter and Consciousness* (1988; Cambridge, MA: MIT Press, 1999). Of course there remain many further interesting philosophical questions about whether such functionalism determines *all* relevant senses of value, meaning, consciousness, ethical worth, and so on. Like many philosophers, I am inclined to say yes—but it is beside the point here.
6. Beliefs are the paradigmatic indicatives, and desires are the paradigmatic imperatives, but there are surely many levels of mental content that fish or mice or robots might have that are not as sophisticated as beliefs and desires. For a better catalog of ways that our representations differ from those of simpler cognitive systems, see the conclusion of Ruth Garrett Millikan, "Biosemantics," in *White Queen Psychology and Other Essays for Alice*. (Cambridge, MA: MIT Press, 1989), 83–101. (I am using "representation" in a broad sense, roughly synonymous with other philosophical terms of art like "intentionality" and "mental content.")
7. See Aristotle, *Nicomachean Ethics* (350 BCE), trans. W. D. Ross, MIT Classics, accessed February, 28, 2020, http://classics.mit.edu/Aristotle/nicomachaen.html, book III; David Hume, *A Treatise of Human Nature* (1739), ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1896), https://books.google.com/books/about/A_Treatise_of_Human_Nature.html?id=5zGpC6mL-MUC, 2.3.3.
8. The standard RL text is Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 1998).
9. (Bostrom loc. 4388).
10. Just in case such short-circuiting sounds at all farfetched, consider that nature designed orgasms to reward reproductive behavior—and that we humans (and many other animals) have found ways to achieve this reward without the intended behavior.
11. The term "thinkful wishing" is from collaboration with Eric Lormand.
12. Bill Hibbard, "Model-Based Utility Functions," *Journal of Artificial General Intelligence* 3, no. 1 (2012): 1–24.
13. Tom Everitt and Marcus Hutter, "Avoiding Wireheading with Value Reinforcement Learning," *arXiv*, May 10, 2016, http://arxiv.org/pdf/1605.03143v1.pdf, 2n1. As a reviewer points out, this applies only in general to utility agents; we *could* design ones whose utility function would enjoin them to enter the experience machine.

14. Related ontological concerns are in Peter De Blanc, "Ontological Crises in Artificial Agents' Value Systems" (San Francisco: Machine Intelligence Research Institute, May 19, 2011), https://intelligence.org/files/OntologicalCrises.pdf.
15. Daniel Dewey, "Learning What to Value" (San Francisco: Machine Intelligence Research Institute, 2011), https://intelligence.org/files/LearningValue.pdf.
16. Where $U_i(w) \in \mathbb{U}$ is a utility function scoring possible worlds, and $v(U_i)$ is the "value criterion" (most generically, "U_i is the correct target utility function"), AI-VL estimates the target utility function and so the value of any possible world as $\hat{U}(w) = \sum_{U_i \in \mathbb{U}} U_i(w) P(v(U_i) \mid w)$.
17. (Bostrom loc. 4564).
18. Everitt and Hutter, "Avoiding Wireheading with Value Reinforcement Learning," propose a value-learning system VRL, a hybrid between utility agent and RL, which learns its utility function through reinforcement. Everitt and Hutter then show that a standard such VRL will have incentive to "optimise its evidence" toward "a more easily satisfied utility function" (10)—in other words, to thinkfully wish. They propose a fix for this concern but rightly worry about its generality.
19. (Bostrom loc. 4473)
20. For the seminal paper, see Andrew Y. Ng and Stuart J. Russell, "Algorithms for Inverse Reinforcement Learning," in *ICML '00*: *Proceedings of the Seventeenth International Conference on Machine Learning* (San Francisco: Morgan Kaufmann, 2000), 663–70, http://dl.acm.org/citation.cfm?id=645529.657801. For a more flexible (and more computationally troublesome) take, see Can Eren Sezener, "Inferring Human Values for Safe AGI Design," in *Artificial General Intelligence*, ed. Jordi Bieger, Ben Goertzel, and Alexey Potapov (Switzerland: Springer International, 2015), 152–55. And for incorporating the observed agent's feedback ("cooperative inverse reinforcement learning"), see Dylan Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning," in *Advances in Neural Information Processing Systems 29*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, 2016), 3909–17, http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf.
21. Aristotle, *Nicomachean Ethics*, 1097b22.
22. Elijah Millgram, "Specificationism," in *Reasoning: Studies of Human Inference and Its Foundations*, ed. Jonathan E. Adler and Lance J. Rips (Cambridge: Cambridge University Press, 2008), 744. For extended treatments, see Aurel Kolnai, "Deliberation Is of Ends," in *Varieties of Practical Reasoning*, ed. Elijah Millgram (Cambridge, MA: MIT Press, 1962), 259–78; Henry S. Richardson, *Practical Reasoning about Final Ends* (Cambridge: Cambridge University Press, 1994).
23. See Eliezer Yudkowsky, "Complex Value Systems Are Required to Realize Valuable Futures" (San Francisco: Machine Intelligence Research Institute, 2011), https://intelligence.org/files/ComplexValues.pdf.
24. Richardson, *Practical Reasoning about Final Ends*, 137.
25. As Elijah Millgram puts it, "Coherence is a vague concept; we should expect it to require specification" ("Specificationism," 741). Note in particular that the coherence

sought here is not (just) the *probabilistic* coherence demanded by Bayesian reasoning, familiar to many AI theorists.

26. Paul Thagard and Karsten Verbeurgt, "Coherence as Constraint Satisfaction," *Cognitive Science* 22, no. 1 (1998): 1–24; Paul Thagard, *Computational Philosophy of Science* (1988; Cambridge, MA: MIT Press, 1993). In collaboration with Millgram, Thagard developed accounts of *deliberative* coherence in Elijah Millgram and Paul Thagard, "Deliberative Coherence," *Synthese* 108, no. 1 (1996): 63–88, and Paul Thagard and Elijah Millgram, "Inference to the Best Plan: A Coherence Theory of Decision," in *Goal-Driven Learning*, ed. Ashwin Ram and David B. Leake (Cambridge, MA: MIT Press, 1995), 439–54; see also Paul Thagard, *Coherence in Thought and Action* (Cambridge, MA: MIT Press, 2000). Though inspired by such work, I now lean toward an alternative Millgram also mentions; see, e.g., Peter D. Grünwald, *The Minimum Description Length Principle* (Cambridge, MA: MIT Press, 2007).
27. Note that Richardson, *Practical Reasoning about Final Ends*, would not agree; see his section 26. (His account relies instead on a "sovereign deliberator" that I find dubious in light of naturalism and AI.)
28. Karl Friston et al., "Active Inference and Epistemic Value," *Cognitive Neuroscience* 6, no. 4 (2015): 187–214.
29. Daniel C. Dennett, "True Believers: The Intentional Strategy and Why It Works" (1981), in *The Intentional Stance* (Cambridge, MA: MIT Press, 1996), 13–35.
30. The case is discussed extensively in Jerry Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990), but is older than that; the source reference tends to be J. Y. Lettvin et al., "What the Frog's Eye Tells the Frog's Brain," *Proceedings of the IRE* 47, no. 11 (1959): 1940–51, https://doi.org/10.1109/JRPROC.1959.287207.
31. Karen Neander, "Teleological Theories of Mental Content," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Metaphysics Research Lab, Stanford University, Spring 2018, https://plato.stanford.edu/archives/spr2018/entries/content-teleological/ has a good overview of the indeterminacy problem in the context of "teleological" theories for reading mental content from physical facts. The best example of such a theory, perhaps, is based in Ruth Garrett Millikan's seminal *Language, Thought, and Other Biological Categories* (1984; Cambridge, MA: MIT Press, 1995).
32. See, e.g., Daniel C. Dennett, "Evolution, Error, and Intentionality" (1987), in *The Intentional Stance* (Cambridge, MA: MIT Press, 1996), 287–321.
33. Dennett, "True Believers," 30–31.
34. At least, our access to and attributions of mental content will be more constrained, if not the content itself.

    Note that Fred Dretske, "Misrepresentation," in *Belief: Form, Content, and Function*, ed. Radu J. Bogdan (Oxford: Oxford University Press, 1986), 17–36, takes the *learning* aspect to be especially important; as long as there are a fixed number of sensory routes $s_1, s_2 \ldots s_n$ to fly detection, we can always say what's *really* meant is "s_1, *or* s_2, *or* $\ldots s_n$" rather than "fly." But not so if the set of perceptual routes is indeterminate, depending on what the creature learns.

35. Honestly I often think of this on the simple model from a computer game I used to play (back before my own reproductive successes), *The Sims*. To keep your simulated person happy in the game requires maintaining several ever-decaying signals at once: "hunger," "social," "bladder," "hygiene," "energy" (requires enough rest), and "fun."
36. I got this story from Thagard, who recently claims pretty good corroboration for it; see, e.g., the opening of chapter 6 in Paul Thagard, *The Brain and the Meaning of Life* (Princeton, NJ: Princeton University Press, 2010).
37. Howard Raiffa and Stephen E. Fienberg, "The Early Statistical Years: 1947–1967: A Conversation with Howard Raiffa," *Statistical Science* 23, no. 1 (2008), 142, http://www.jstor.org/stable/27645884.
38. These are not meant as actual examples of what I mean by "simple" values in humans, which I take ultimately to be biological, fixed reinforcers roughly like the "four Fs" (food, fight, flight, and reproduction). Thus a *relatively* simple value like "adventure" might itself be a complex blend of lower-level reinforcers to do with novelty and how it is registered in the brain (biologically as dopamine, or computationally as surprisal measure, etc.).
39. For a nuanced discussion of such commensurability, see chapter 6 of Richardson, *Practical Reasoning about Final Ends*.
40. See Weija Wang, "Multi-Objective Sequential Decision Making" (PhD diss., Université Paris Sud-Paris XI, 2014); Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári, "Multi-Criteria Reinforcement Learning," in *ICML '98*: *Proceedings of the Fifteenth International Conference on Machine Learning* (San Francisco: Morgan Kaufmann, 1998), 197–205, http://dl.acm.org/citation.cfm?id=645527.657298. Another strategy besides scalarizing is to treat each Pareto-optimal policy proposal as a kind of subagent with negotiating power; see Andrew Critch, "Toward Negotiable Reinforcement Learning: Shifting Priorities in Pareto Optimal Sequential Decision-Making," *arXiv*, last revised May 13, 2017, http://arxiv.org/abs/1701.01302. I might mention that yet another type of approach to reconciling multiple basic values is to elaborate the DECO model of deliberative coherence from Millgram and Thagard, "Deliberative Coherence," into a model of "belief-desire coherence"—as I have previously sought to do: Steve Petersen, "Belief-Desire Coherence" (PhD diss., University of Michigan, 2003).
41. This is oversimplified in part because an intelligent agent would *learn* some of these as instrumental goals.
42. I mean the reward signals or utility functions can be "simple" in the sense of low Kolmogorov complexity: essentially, they require relatively few lines of code to specify precisely. Calculating the combined total is "complex" in the different sense that, as in other weighted constraint satisfaction problems, finding the vector to optimize the scalar typically cannot be done in reasonable amounts of time (even by a superintelligence), and must be approximated.
43. There is more to be said about when and whether there *is* an "underlying phenomenon." I will not be saying it here, though.
44. Yudkowsky, "Complex Value Systems Are Required to Realize Valuable Futures," p. 3.

45. The seminal statement is in John Rawls, *A Theory of Justice* (1971; Cambridge, MA: Harvard University Press, 1995), with elaboration in, e.g., Norman Daniels, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76, no. 5 (1979): 256–82.
46. Reflective equilibrium over human value judgments seems as though it would result in something closely related to the "coherent extrapolated volition" from Eliezer Yudkowsky, "Coherent Extrapolated Volition" (San Francisco: Machine Intelligence Research Institute, 2004), https://intelligence.org/files/CEV.pdf; Bostrom discusses the proposal in some detail starting from loc. 4907.
47. Thanks to Einar Duenger Bøhn, John Danaher, Matthew Liao, Eric Schwitzgebel, Marija Slavkovik, and two anonymous reviewers.

References

Andersen, Ross. 2013. "Omens." *Aeon*, 2013. https://aeon.co/essays/will-humans-be-around-in-a-billion-years-or-a-trillion.

Aristotle. *Nicomachean Ethics* (350 BCE). Translated by W. D. Ross. MIT Classics. Accessed February 28, 2020. http://classics.mit.edu/Aristotle/nicomachaen.html.

Bostrom, Nick. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 12–17. Windsor ON: International Institute for Advanced Studies in Systems Research / Cybernetics, 2003.

Churchland, Paul M. *Matter and Consciousness*. 1988; Cambridge, MA: MIT Press, 1999 edition.

Critch, Andrew. "Toward Negotiable Reinforcement Learning: Shifting Priorities in Pareto Optimal Sequential Decision-Making." *arXiv*, last revised May 13, 2017. http://arxiv.org/abs/1701.01302.

Daniels, Norman. "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy* 76, no. 5 (1979): 256–82.

De Blanc, Peter. "Ontological Crises in Artificial Agents' Value Systems." San Francisco: Machine Intelligence Research Institute, May 19, 2011. https://intelligence.org/files/OntologicalCrises.pdf.

Dennett, Daniel C. "Evolution, Error, and Intentionality" (1987). In *The Intentional Stance*, 287–321. Cambridge, MA: MIT Press, 1996.

Dennett, Daniel C. "True Believers: The Intentional Strategy and Why It Works" (1981). In *The Intentional Stance*, 13–35. Cambridge, MA: MIT Press, 1996.

Dewey, Daniel. "Learning What to Value." San Francisco: Machine Intelligence Research Institute, 2011. https://intelligence.org/files/LearningValue.pdf.

Dickens, Charles. *A Christmas Carol in Prose: Being a Ghost Story of Christmas*. London: Chapman & Hall, 1843. Project Gutenberg. http://www.gutenberg.org/ebooks/46.

Dretske, Fred. "Misrepresentation." In *Belief: Form, Content, and Function*, edited by Radu J. Bogdan, 17–36. Oxford: Oxford University Press, 1986.

Everitt, Tom, and Marcus Hutter. "Avoiding Wireheading with Value Reinforcement Learning." *arXiv*, May 10, 2016. http://arxiv.org/pdf/1605.03143v1.pdf.

Fodor, Jerry. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press, 1990.

Friston, Karl, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. "Active Inference and Epistemic Value." *Cognitive Neuroscience* 6, no. 4 (2015): 187–214.

Gábor, Zoltán, Zsolt Kalmár, and Csaba Szepesvári. "Multi-Criteria Reinforcement Learning." In *ICML '98*: *Proceedings of the Fifteenth International Conference on Machine Learning*, 197–205. San Francisco: Morgan Kaufmann, 1998. http://dl.acm.org/citation.cfm?id=645527.657298.

Grünwald, Peter D. *The Minimum Description Length Principle*. Cambridge, MA: MIT Press, 2007.

Hadfield-Menell, Dylan, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. "Cooperative Inverse Reinforcement Learning." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 3909–17. Red Hook, NY: Curran Associates, 2016. http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf.

Hibbard, Bill. "Model-Based Utility Functions." *Journal of Artificial General Intelligence* 3, no. 1 (2012): 1–24.

Hume, David. *A Treatise of Human Nature* (1739). Edited by L. A. Selby-Bigge. Oxford: Clarendon Press, 1896. https://books.google.com/books/about/A_Treatise_of_Human_Nature.html?id=5zGpC6mL-MUC.

Kolnai, Aurel. "Deliberation Is of Ends." In *Varieties of Practical Reasoning*, edited by Elijah Millgram, 259–78. Cambridge, MA: MIT Press, 1962.

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. "What the Frog's Eye Tells the Frog's Brain." *Proceedings of the IRE* 47, no. 11 (1959): 1940–51. https://doi.org/10.1109/JRPROC.1959.287207.

Millgram, Elijah. "Specificationism." In *Reasoning: Studies of Human Inference and Its Foundations*, edited by Jonathan E. Adler and Lance J. Rips, 731–47. Cambridge: Cambridge University Press, 2008.

Millgram, Elijah, and Paul Thagard. "Deliberative Coherence." *Synthese* 108, no. 1 (1996): 63–88.

Millikan, Ruth Garrett. "Biosemantics." In *White Queen Psychology and Other Essays for Alice*, 83–101. Cambridge, MA: MIT Press, 1989.

Millikan, Ruth Garrett. *Language, Thought, and Other Biological Categories*. 1984; Cambridge, MA: MIT Press, 1995.

Neander, Karen. "Teleological Theories of Mental Content." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, Spring 2018. https://plato.stanford.edu/archives/spr2018/entries/content-teleological/.

Ng, Andrew Y., and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning." In *ICML '00*: *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–70. ICML '00. San Francisco: Morgan Kaufmann, 2000. http://dl.acm.org/citation.cfm?id=645529.657801.

Nozick, Robert. *Anarchy, State, and Utopia*. Oxford: Basil Blackwell, 1974.

Petersen, Steve. "Belief-Desire Coherence." PhD diss., University of Michigan, 2003.

Petersen, Steve. "Superintelligence as Superethical." In *Robot Ethics 2.0*, edited by Patrick Lin, Ryan Jenkins, and Keith Abney, 322–37. New York: Oxford University Press, 2017.

Raiffa, Howard, and Stephen E. Fienberg. "The Early Statistical Years: 1947-1967. A Conversation with Howard Raiffa." *Statistical Science* 23, no. 1 (2008): 136–49. http://www.jstor.org/stable/27645884.

Rawls, John. *A Theory of Justice*. 1971; Cambridge, MA: Harvard University Press, 1995.

Richardson, Henry S. *Practical Reasoning about Final Ends*. Cambridge: Cambridge University Press, 1994.

Sezener, Can Eren. "Inferring Human Values for Safe AGI Design." In *Artificial General Intelligence*, edited by Jordi Bieger, Ben Goertzel, and Alexey Potapov, 152–55. Cham, Switzerland: Springer International, 2015.

Shah, Rohin. "Value Learning." *AI Alignment Forum*, October 29, 2018. https://www.alignmentforum.org/s/4dHMdK5TLN6xcqtyc.

Soares, Nate. "The Value Learning Problem." San Francisco: Machine Intelligence Research Institute, 2016. https://intelligence.org/files/ValueLearningProblem.pdf.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

Thagard, Paul. *The Brain and the Meaning of Life*. Princeton, NJ: Princeton University Press, 2010.

Thagard, Paul. *Coherence in Thought and Action*. Cambridge, MA: MIT Press, 2000.

Thagard, Paul. *Computational Philosophy of Science*. 1988; Cambridge, MA: MIT Press, 1993.

Thagard, Paul, and Elijah Millgram. "Inference to the Best Plan: A Coherence Theory of Decision." In *Goal-Driven Learning*, edited by Ashwin Ram and David B. Leake, 439–54. Cambridge, MA: MIT Press, 1995.

Thagard, Paul, and Karsten Verbeurgt. "Coherence as Constraint Satisfaction." *Cognitive Science* 22, no. 1 (1998): 1–24.

Wang, Weija. "Multi-Objective Sequential Decision Making." PhD diss., Université Paris Sud-Paris XI, 2014.

Yudkowsky, Eliezer. "Coherent Extrapolated Volition." San Francisco: Machine Intelligence Research Institute, 2004. https://intelligence.org/files/CEV.pdf.

Yudkowsky, Eliezer. "Complex Value Systems Are Required to Realize Valuable Futures." San Francisco: Machine Intelligence Research Institute, 2011. https://intelligence.org/files/ComplexValues.pdf.

Part IV: Artificial Intelligence, Consciousness, and Moral Status

15

How to Catch an AI Zombie

Testing for Consciousness in Machines

Susan Schneider

How can we determine if AI is conscious? In projects at the Institute for Advanced Study in Princeton and YHouse in New York, Edwin Turner and I are designing tests for machine consciousness. Our starting point is simple. We do not need to await agreement on a formal definition of consciousness, nor do we require a complete understanding of the neural basis of consciousness, to ask about machine consciousness. For each of us can grasp what is essential about consciousness. Every moment of your waking life, and whenever you dream, it feels like something to be you. When you see the rich hues of a sunrise or smell the aroma of your morning coffee, you are having conscious experience. This is what we mean by "consciousness."

In this chapter, I sketch a provisional framework for investigating artificial consciousness. According to this framework, detecting machine consciousness is like diagnosing a medical illness. There may be a variety of useful tests and markers for detecting synthetic consciousness, some of which are more authoritative than others, and the appropriate use of any one test depends upon contextual factors. Further, because the tests are at an early stage of development, where two or more tests can be used, the results should be checked against each other, in hopes that the tests will themselves be refined and that new tests will be created. The tests for machine consciousness I discuss are the following: the AI Consciousness Test (ACT) (developed with Edwin Turner), my Chip Test, and a test based on the Integrated Information Theory (IIT), developed by Giulio Tononi and others.

This chapter is organized as follows. Section 15.1 stresses the importance of developing tests for synthetic consciousness. Section 15.2 explores the ACT. In Section 15.3, I venture some general methodological remarks. Section 15.4 offers a brief discussion of IIT. Section 15.5 turns to my Chip Test. Finally, Section 15.6 concludes by considering certain pragmatic considerations.

Susan Schneider, *How to Catch an AI Zombie* In: *Ethics of Artificial Intelligence.* Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905040.003.0016

15.1. What Is at Stake?

The science fiction–like flavor of the topic of AI consciousness shouldn't distract us from the very real possibility that certain future AIs might, at some point, be conscious. Robots are already being built to tug at our heartstrings, such as the life-like androids at both Hanson Robotics and Hiroshi Ishugaru's lab. The public will inevitably suspect that such robots with appearance-based bells and whistles have feelings. And as AI grows in sophistication, beyond domain-specific systems that merely excel at games like Go and *Jeopardy!* to AIs with sophisticated, domain-general capacities, a case can be made that certain AIs may be conscious. I'll now outline five reasons why it is important to begin developing tests for machine consciousness, even now.

1. *Tests for AI consciousness and the nature of mind.* Notice that in the biological domain, the more intelligent organisms seem to have richer mental lives. However, this generalization may be incorrect when applied to AIs. As my project with NASA suggests, for all we know, the greatest intelligences in the universe may be postbiological, being superintelligent AIs that are not conscious.[1] Nonconscious AIs would neither be selves nor have minds, for it wouldn't feel like anything to be them. This would mean, intriguingly, that sophisticated intelligence and mindedness do not go hand in hand, as in the biological arena.
2. *Conscious AIs and moral consideration.* By developing synthetic intelligence, humans are entering the era of *intelligent design*, in which we humans, not God, are the designers. Robots are already being designed to take care of the elderly in Japan, be our personal assistants, clean up nuclear reactors, fight our wars, and more. Naturally the question has arisen: Would it be ethical to use a conscious AI for these sorts of tasks? While few would claim that today's AIs are conscious, there seems to be general agreement that conscious AI, should it ever exist, deserves special moral consideration, alongside other conscious beings. To avoid intentional or inadvertent harm to conscious beings, it is important that we know what we are building and developing tests for consciousness in machines.
3. *A human-machine merger.* Silicon-based neural prosthetics are currently being developed, or have been developed, to treat patients with depression, Parkinson's disease, memory loss, and posttraumatic stress disorder. Two well-known companies, Neuralink and Kernel, as well as a large program at the Defense Advanced Research Projects Agency (DARPA), have as their explicit aim to design enhancements to merge humans and machines. Suppose this trend toward AI-based neural prosthetics and enhancements continues, and it is 2060 and you have learned you have a

brain tumor and only have a few months to live. At your doctor's recommendation, you visit a new facility, iBrain, where the surgeon tells you that each part of your brain can be replaced by microchips, one after the next, until, over the course of a few hours, you emerge from surgery with an entirely artificial brain.

Should you do it? Notice that if machines cannot be conscious, then humans who attempt to merge with AI by transferring their minds to a synthetic substrate will not be conscious either. AI-based neural prosthetics will not support consciousness. So you would, in effect, lose the last few months of your life as a conscious being. More generally, all the biological beings who fully merge with AIs actually become nonconscious simulacra of themselves. They will become what philosophers call "AI zombies."

4. *Being supplanted by nonconscious machines.* Elon Musk, Stephen Hawking, Nick Bostrom, Bill Gates, and many others have expressed the concern that humans could quickly lose control of superintelligent AIs, leading to human extinction. This has been called "the control problem." Bostrom stresses that this need not be due to any malevolence on the AI's part; the AI may merely have goals that inadvertently lead to human extinction.[2] In a similar vein, suppose that the superintelligence that supplants humans has some sort of existence that humans can identify with (e.g., intellectual curiosity, the creation of art, etc.) rather than devoting all its time to banal tasks like making paper clips, as in Bostrom's well-known example. Suppose further that it seems to have a rich inner world, with varied sensory experiences, immense attentional and reflective abilities, and so on. In this case, some readers might take some consolation in the fact that the extinction of *Homo sapiens* was at least followed by the presence of a greater intelligence that realized forms of consciousness that we can appreciate. But now assume, instead, that the superintelligence that supplants humans is not conscious. This would seem to be even more tragic, as no one except nonhuman animals would remain to experience the world.

The scenarios depicted in (3) and (4) concern existential risks involving the replacement of conscious humans by nonconscious machines, representing a perverse instantiation of AI technology. The very technology that is supposed to make life easier for humans actually leads to their demise. Ironically, Elon Musk has recently suggested that humans merge with AIs so humans can avoid being outmoded by AIs in the workplace and to remain on same intellectual plane as superintelligence. But as scenario (3) indicates, if AI is not conscious, a merger with AI will hardly promote human flourishing, at least if the so called mind-machine merger replaces the parts of the brain underlying consciousness with

AI technology. This process would instead replace conscious humans with nonconscious simulacra.

5. *Synthetic consciousness and our ability to control AI.* It is currently a matter of debate whether consciousness could make a given type of AI system more safe, less safe, or have no impact whatsoever. On the one hand, perhaps consciousness could make a machine more volatile, as having intense feelings could make it more prone to act in unforeseen ways, like HAL 9000 in Stanley Kubrick's *2001: A Space Odyssey.* On the other hand, machine consciousness could do the opposite: using its own subjective experience as a springboard, superintelligent AI could recognize in humans the capacity for conscious experience and be more empathetic. After all, to the extent that humans value the lives of nonhuman animals, we value them because we believe they can suffer and feel a range of emotions. If this is the case, developing machine consciousness could help solve the control problem. These considerations suggest that we need tests for synthetic consciousness so that the impact consciousness has on AI safety can be assessed.[3]

These considerations illustrate that there are potentially very serious real-world costs to getting facts about AI consciousness wrong. It is imperative to sharpen our understanding of machine consciousness. So, how can we determine whether a given AI is conscious? I'll now propose a rough framework.

15.2. Testing AI Consciousness: The ACT

You might think that we could simply look under the hood, examining the architecture of the AI. But even today programmers are having difficulties understanding why deep-learning systems do what they do. (This is called the "Black Box Problem.") Further, even if a map of the cognitive and perceptual architecture of a sophisticated AI were laid out in front of us, how would we recognize certain architectural features as those central to consciousness? An AI has no brainstem, no claustrum. It is only by analogy with ourselves that we come to believe nonhuman animals are conscious: they have nervous systems and brains; machines do not. Further, even if we think we have a handle on an AI's inner workings at one moment, its design can quickly morph into something too complex for human understanding.

What if the AI's architecture contains cognitive functions like our own, including those correlated with consciousness in the biological case, such as attention and working memory? While these features are suggestive of consciousness, consciousness may also depend upon low-level details specific to the type of

material out of which the AI is constructed. As I've stressed elsewhere, the properties that an AI needs to complete a task of financial value to an AI company may not be the same properties that give rise to consciousness. The low-level details could matter.[4] Bearing all this in mind, Edwin Turner and I have suggested that the following behavior-based test for synthetic consciousness could be useful.
Notice that normal adults can *quickly* and *readily* grasp certain concepts based on the quality of felt consciousness—that is, the way it feels, from the inside, to experience the world.[5] Consider, for instance, the film *Freaky Friday*, in which a mother and daughter switch bodies with each other. Filmgoers found this scenario imaginable because, as conscious beings, they can imagine their mind being in an entirely different body. In a similar vein, we can at least roughly consider the possibility of an afterlife, of being reincarnated, or of having an out-of-body experience. The point is not that we believe such scenarios to be true; the point is we can imagine them, at least in broad stokes, because we are conscious beings.

These scenarios would be exceedingly difficult to comprehend for an entity that had no conscious experience whatsoever. It would be like expecting someone who is completely deaf from birth to appreciate a Beethoven symphony. This simple observation leads to a test for AI consciousness that singles out AIs with phenomenal consciousness from those that lack phenomenal consciousness. For it is the inner feeling of our mental lives that allows one to imagine these scenarios.

An ACT would challenge an AI with a series of increasingly demanding natural-language interactions to see how *quickly* and *readily* it can grasp and use concepts based on the internal experiences we associate with consciousness. The test would have multiple questions; a satisfactory answer to any question would be sufficient for passing the test. At the most elementary level we might simply ask the machine if it conceives of itself as anything other than its physical self. We could see how the AI deals with ideas and scenarios such as reincarnation, out-of-body experiences, body switching, and so on. We might also run a series of experiments to see whether the AI tends to prefer certain kinds of events to occur in the future as opposed to the past. A nonconscious AI should have no preference whatsoever. If there appears to be a preference, we should ask the AI to explain its answer, if it has linguistic abilities. Conscious beings are in the experienced present, but our subjective sense presses onto the future. We generally wish for positive experiences in the future and wish to minimize painful ones. But from the vantage point of the physical laws, there is no reason that a system should prefer the future to the past, as both General Relativity Theory and the Standard Model of particle physics are temporally symmetric. These laws do not say whether time is moving forward or backward. Nor do the laws identify so special a moment as appears to us as what we call "now."[6] We might also see if the

AI seeks out alternative states of consciousness, for example, when given the opportunity to modify its own weights or parameters in radical ways or somehow inject noise into the system.

An AI may even illustrate a sophisticated understanding of consciousness. To probe for this, we might ask it more demanding questions. (It should be noted, however, that having a sophisticated understanding of conscious experience does not imply that a system is somehow "more conscious" than another that responds only to less demanding questions. I make no comparative claims here.) At an advanced level, an AI's ability to reason about and discuss philosophical issues such as the hard problem of consciousness, the mind-body problem, zombie cases, and the problem of spectrum inversion would be evaluated. At the most demanding level, we might see if the machine invents and uses consciousness-based concepts on its own, without our prompts. Perhaps it is curious about whether we are conscious, despite the fact that we are biological!

The following example illustrates the general idea. Suppose we find a planet that has a highly sophisticated, silicon-based life form (call them the "Zetas"). Scientists observing the Zetas begin to ask whether they are conscious. What would be convincing proof of their consciousness? If the Zetas express curiosity about whether there is an afterlife or ponder whether they are more than just their bodies, it would be reasonable to judge them as conscious. There are also nonverbal cultural behaviors that could indicate Zeta consciousness, such as mourning the dead (which indicates a sense that the deceased being has a mind or is a self), religious activities, or even turning colors in situations that correlate with emotional challenges, as chromatophores do on Earth. Such behaviors could indicate that it feels like something to be a Zeta.

The death of the fictional HAL 9000 in the film *2001: A Space Odyssey* is another example. HAL neither looks nor sounds like a human being. (A human did supply HAL's voice, but in an eerie, flat way.) Nevertheless the *content* of what HAL says as it is deactivated by the astronaut—HAL pleads with the astronaut to spare it from impending "death"—conveys a powerful impression that HAL is a conscious being.

Could these sorts of behaviors help to identify conscious AIs on Earth? Here an obstacle arises. Even today's robots can be programmed to make convincing utterances about consciousness, and a highly intelligent machine could perhaps even use information about neurophysiology to infer the presence of consciousness in biological creatures. If sophisticated nonconscious AIs aim, for whatever reason, to mislead us into believing that they are conscious, their knowledge of human consciousness and neurophysiology could help them do so.

We can get around this, though. One proposed technique in AI safety involves "boxing in" an AI—making it unable to get information about the outside world or act outside of a circumscribed domain, that is, the "box." To box in an AI for

the purpose of conducting an ACT, the AI should not have access to the internet, where it could learn about neurophysiology, phenomenal consciousness, and so on. Nor should it have access to literary or academic works introducing these themes. The AI could still have natural-language abilities, however. Learning a vocabulary that includes expressions like "believes," "you," and "perspective" need not be prohibited. The use of such expressions, in and of themselves, does not cause us to believe a machine is conscious. For instance, consider that IBM's Watson used such expressions during the *Jeopardy!* matches with an impressive degree of competence, yet people did not come to believe it was conscious.[7]

Would an AI that is boxed in be unable to communicate to us that it is conscious, since it lacked information about consciousness and the brain? Because a conscious being's primary sense of consciousness comes from first-person experience, it seems plausible that clever testing strategies could enable one to convey first-person experience to others. Here, Isaac Asimov's tale "Robot Dreams" is illustrative. The story begins with the robot, LVX-1 (called "Elvex") stating to its amazed creators, baldly, "Last night I dreamed." It is worth reproducing the dialogue:

> The robot's head turned toward her smoothly. "Yes, Dr. Calvin?"
>
> "How do you know you have dreamed?"
>
> "It is at night, when it is dark, Dr. Calvin," said Elvex, "and there is suddenly light, although I can see no cause for the appearance of light. I see things that have no connection with what I conceive of as reality. I hear things. I react oddly. In searching my vocabulary for words to express what was happening, I came across the word 'dream.' Studying its meaning I finally came to the conclusion I was dreaming."
>
> "How did you come to have 'dream' in your vocabulary, I wonder."
>
> Linda said, quickly, waving the robot silent, "I gave him a human-style vocabulary. I thought—"
>
> "You really thought," said Calvin. "I'm amazed."
>
> "I thought he would need the verb. You know, 'I never dreamed that—' Something like that."
>
> Calvin said, "How often have you dreamed, Elvex?"
>
> "Every night, Dr. Calvin, since I have become aware of my existence."
>
> "Ten nights," interposed Linda, anxiously, "but Elvex only told me of it this morning."
>
> "Why only this morning, Elvex?"
>
> "It was not until this morning, Dr. Calvin, that I was convinced that I was dreaming. Till then, I had thought there was a flaw in my positronic brain pattern, but I could not find one. Finally, I decided it was a dream."[8]

This is an intriguing example, although, admittedly, Elvex knows more than a boxed-in AI would, as Elvex has information about dreams in its database. A boxed-in AI would be unable to offer the conclusion that it was dreaming, but it could report what it "saw" at night. The job of the test is to generate conclusions; the machine need not conclude that it is conscious or use expressions like "conscious." A boxed-in system could report that it is curious about its internal states, respond to scenarios and questions, and report anomalous findings using descriptions of its internal states. It has the linguistic resources to do this.

ACT could be useful for consciousness engineering during the development of different kinds of AIs, helping to avoid using conscious machines in unethical ways or to create artificial consciousness when appropriate. If a machine passes ACT, other features of the system can then be measured, to see if the presence of consciousness is correlated with higher or lower levels of empathy, volatility, goal content integrity, intelligence, and so on. Other nonconscious versions of the system could serve as a basis for comparison, if available.

Some doubt that a superintelligent machine could be boxed in effectively because a superintelligence could inevitably find a clever escape. We do not anticipate the development of superintelligence over the next two decades, however. We merely hope to provide a method to test some kinds of AIs, not all AIs. Furthermore, for an ACT to be effective, the AI need not stay in the box for long, just long enough for someone to administer the test. So perhaps the test can be administered to early superintelligences, although a superintelligence could, in principle, alter or remove its own consciousness at some later point, perhaps in the process of creating more efficient future versions of itself.[9]

A version of ACT could examine the behavior of a group of AIs. For instance, we might construct an artificial-life program in which several AIs of a certain type evolve. We could look at the evolution of certain linguistic and/or nonlinguistic behaviors, to see if any behaviors suggest consciousness. Behavior alone can be suggestive; consider cases in which elephants, chimpanzees, or certain marine mammals, like Orcas, mourn the dead.[10] Different versions of an A-life program could be tested, each of which has AIs with slightly different capacities, to determine whether and when consciousness evolves.

Different versions of the ACT could be generated, depending upon the context. For instance, one version could apply to nonlinguistic agents within an A-life program, looking for specific behaviors. Another could apply to an AGI with highly linguistic abilities and probe it for sensitivity to religious, body-swapping, or philosophical scenarios involving consciousness. If an ACT contains various probes or questions, a positive indication on any probe or question is sufficient for being judged to be conscious.

ACT resembles Alan Turing's celebrated test for intelligence because it is entirely based on behavior—and, like Turing's, it could be implemented in a

formalized question-and-answer format. But an ACT is also quite unlike the Turing test, which was intended to bypass any need to know what was transpiring inside the "mind" of the machine. By contrast, an ACT is intended to do *exactly the opposite*; it seeks to reveal a subtle and elusive property of the machine's mind. Indeed a machine might fail the Turing test because it cannot pass for a human, but pass an ACT because it exhibits behavioral indicators of consciousness.

This, then, is the underlying basis of our ACT. It is worth reiterating the strengths and limitations of the test. In a positive vein, we believe passing the test is *sufficient* for being conscious; that is, if a system passes it, it can be regarded as phenomenally conscious.[11] So the test is a zombie filter: creatures merely having functional correlates of consciousness (e.g., attention), creativity, or high general intelligence shouldn't pass ACT, at least if they are boxed in effectively (see later discussion of functional correlates of consciousness). ACT filters out zombies by finding *only* those creatures sensitive to the felt quality of experience.

But it may not find *all* conscious AIs. First, an AI could lack the linguistic or conceptual ability to pass a linguistic version of the ACT, like an infant or certain nonhuman animals, yet still be capable of experience. Second, ACT is obviously derived from our own conception of consciousness, and the questions and situations employed rely upon our understanding of phenomenal consciousness (e.g., we can view the mind as potentially separable from the body, we mourn the death of loved ones, and so on). I happen to suspect that at least some of these features would be shared across a range of highly intelligent conscious beings, but it is best to assume that not all highly intelligent conscious beings have such a conception. For these reasons, the ACT should not be construed as a necessary condition that all AIs must pass. Put another way, failing ACT does not mean that a system is not conscious.

I'll now attempt to use ACT, together with other tests, to sketch a provisional framework for AI consciousness.

15.3. Toward a Provisional Framework: Cautionary Remarks

I've stressed that a version of ACT could be used for a range of cases, but candidates for the test must be selected with care. It may be that ACT identifies only a small portion of the larger space of conscious AIs (if such AIs even exist). Perhaps the AIs that ACT identifies as conscious help us in identifying other conscious AIs, ones that are perhaps somehow more alien or inscrutable.

There are further issues to tread on carefully as well. Claims about a species, individual, or AI reaching "heightened levels of consciousness" or a "richer consciousness" should be carefully explained, for they may be implicitly evaluative,

threatening to be "speciesist," to borrow an expression that Singer used in the context of the animal liberation debate. There are a variety of phenomena that such expressions could refer to, and our judgments about this issue are inevitably influenced by our evolutionary history and biology, and they can even be biased by the cultural and economic backgrounds of the academics working on this topic. By such expressions one might mean altered states of consciousness, such as the meditative awareness of a Buddhist monk. Or one could have in mind the consciousness of a creature that has numerous states under the spotlight of attention that somehow feel vivid to it. Alternately, one could be referring to a situation in which a creature has a great number of conscious states, a more varied range of sensory contents, has states that are felt with heightened emotional intensity, or has states that are somehow regarded by us as being more intrinsically valuable (e.g., listening to Bach versus getting drunk), and so on. The tests suggested herein are not meant to establish a hierarchy of experiences, thankfully. They are merely an initial step toward the identification of conscious AIs.

Further, specialists on AI consciousness often distinguish consciousness from an important related notion. The felt quality of one's inner experience—what it feels like, from the inside, to be you—is often called "phenomenal consciousness" (PC) by philosophers. (Herein, I've simply called it "consciousness.") Experts on machine consciousness tend to distinguish PC from what they call "cognitive consciousness" (CC).[12] An AI has CC when it has architectural features that are at least roughly like those found to underlie PC in humans, such as attention and working memory. (Unlike isomorphs, cases of CC need not be precise computational duplicates. They can have simplified versions of human cognitive functions.)

Many do not like to call cognitive consciousness a kind of consciousness at all, for a system with CC, without PC, would be a rather sterile form of consciousness, lacking any subjective experience. Such a system would be an AI zombie. Systems merely having CC may not behave as phenomenally conscious systems do, nor would it be fit to treat these systems as sentient beings. Such systems would not grasp the painfulness of a pain or the warmth of the summer sun.

So why is CC interesting, for our purposes? It is important for at least two reasons. First, perhaps CC is necessary to have the kind of PC that biological beings have. If one is interested in developing conscious machines, this could be important, for if we develop CC in machines, perhaps we would get closer to developing PC in AIs. Second and relatedly, the presence of CC is reasonably regarded as being a marker for the possible presence of PC, and if possible, tests should be carried out. This highlights the import of having a test for PC that singles out AIs with PC from AIs that are nonconscious zombies, merely having features of CC.

The test just discussed, as well as my Chip Test, which is to be discussed shortly, are intended to complement an influential existing approach to machine consciousness, the Integrated Information Theory (IIT).

15.4. The Integrated Information Theory

IIT has been discussed extensively in the literature, so my comments will be brief. IIT was developed by the neuroscientist Giulio Tononi and his collaborators at the University of Wisconsin, Madison. Intrigued by the hard problem of consciousness, Tononi's point of departure is the felt quality of experience. He claims that felt-quality consciousness requires a high level of "integrated information" within a system. Information is "integrated" within a system when the system's states are highly interdependent, featuring a rich web of feedback among its parts.[13] The level of integrated information can in principle be measured (designated by the Greek letter Φ). IIT holds that if we know the value of Φ, we can determine if a system is conscious and how conscious it is.

Could the resources of IIT yield a feasible test for synthetic consciousness, identifying machines that have the requisite Φ level as conscious? Like the ACT, IIT looks beyond superficial features of an AI, such as its humanlike appearance. Indeed different kinds of AI architectures can be compared in terms of their measure of integrated information. The presence of a quantitative measure for phenomenal consciousness would be incredibly useful. Unfortunately the calculations involved in computing Φ for even a small part of the brain, such as the claustrum, are computationally intractable. (That is, Φ can't be calculated precisely except for extremely simple systems.) Simpler metrics that approximate Φ have been provided, however, and the results are encouraging. For instance, the cerebellum has a relatively low Φ level, predicting that it contributes little to the overall consciousness of the brain. This fits with the data. Humans born without a cerebellum (a condition called "cerebellar agenesis") do not seem to differ from normal subjects in the level and quality of their conscious lives. The cerebellum has low interconnectedness, exhibiting feedforward processing. In contrast, parts of the brain that, when injured or missing, contribute to a certain kind of loss in conscious experience have higher Φ values. IIT is also able to discriminate between levels of consciousness in normal humans (wakefulness versus sleep) and even single out "locked in" patents, who are unable to communicate.[14]

IIT is what astrobiologists call a "small n" approach, that is, an approach that reasons from the biological case on Earth to a broader range of cases (in this case, the class of conscious machines). This is an understandable drawback, however, as the biological case is the only case of consciousness we know of. The tests I propose also have this drawback. Biological consciousness is the only case we

know of, so we had better use it as our point of departure, together with a heavy dose of humility.

Another feature of IIT is that it ascribes a small amount of consciousness to anything that has a minimal amount of Φ. In a sense, this is akin to the doctrine of panpsychism, as microscopic and inanimate objects could have at least a small amount of experience. But this does not mean that the view is panpsychist, at least if panpsychism is construed as claiming that everything has at least a small amount of experience. For IIT does not ascribe consciousness to everything. In fact IIT does not predict that feedforward computational networks are conscious, for they lack sufficient causal integration between the components. As Tononi and Koch note, "ITT predicts that consciousness is graded, is common among biological organisms and can occur in some very simple systems. Conversely, it predicts that feed-forward networks, even complex ones, are not conscious, nor are aggregates such as groups of individuals or heaps of sand."[15]

IIT singles out certain systems as conscious in a special sense, however. That is, it aims to predict which systems have a more complex form of consciousness, akin to that what occurs in normally functioning brains.[16] The question of AI consciousness, in this context, seeks to determine whether machines have macroconsciousness as opposed to smaller Φ levels exhibited by everyday objects.

Is having high Φ sufficient for a machine's being conscious? According to Scott Aaronson, the director of the Quantum Information Center at the University of Texas at Austin, a two-dimensional grid that runs error-correction codes such as those used for CDs will have a very high Φ level. Aaronson writes, "IIT predicts not merely that these systems are 'slightly' conscious (which would be fine) but that they can be unboundedly more conscious than humans are."[17] But a grid does not seem to be the sort of thing that is conscious, suggesting to many that IIT should not be regarded as being sufficient for consciousness.

Tononi has responded to Aaronson's point by biting the bullet, asserting that the grid is conscious (i.e., macroconscious)! I prefer to instead reject the view that having a high Φ value is *sufficient* for an AI to be conscious. But IIT could supply a necessary feature that all conscious systems have. For all we know, all conscious systems, biological or mechanical, may have the requisite minimum level of Φ.

How are we to deal with a machine that has high Φ, should we ever encounter one? We've seen that Φ is probably not sufficient. Further, since research on Φ has been in biological systems and today's computers are not good candidates for being conscious, it is too early to tell whether Φ is a necessary condition for AI consciousness. In the final section of the paper, I suggest a way of dealing with this situation. But before we delve into this, it will be helpful to have the third test on the table.

15.5. The Chip Test

Now let's turn to a different sort of test. Silicon-based brain chips are already under development as a treatment for various memory-related conditions, such as Alzheimer's and PTSD, and companies like Kernel and Neuralink aim to develop AI-based brain enhancements for healthy individuals. In a similar vein, consider the following hypothetical scenario.

Suppose it is 2045, and you have just learned you have an extensive brain tumor. You go to a center called iBrain, where researchers are working to gradually replace parts of the brain with brand-new, durable microchips. You agree to the surgery because you are desperate for a cure, but you are aware that the prosthetics they use may not be perfect functional duplicates of the original parts of the brain they replace, as the science of neural prosthetics is not perfected at this time.

During the surgery you are to remain awake, and you will need to report any changes to the felt quality of your consciousness. The surgeons need to replace various parts of the brain that are central to consciousness, and they are especially keen to learn whether any aspect of your consciousness is impaired when they replace your brain tissue with the neural prosthetics. Their hope is that as the technology is perfected, they will be able to use perfect neural prostheses in areas of the brain underlying consciousness without any change in the quality of your conscious experience. But they are not sure whether their prosthetics will work—they will watch and wait.

If, during this process, a prosthetic part of the brain ceases to function normally—specifically, if it ceases to give rise to the aspect of consciousness that that brain area is responsible for—then there should be behavioral indications, including verbal reports. An otherwise normal person should be able to detect, or at least indicate to others through odd behaviors, that something is amiss, as with traumatic brain injuries involving the loss of consciousness in some domain.

This would indicate a "substitution failure" of the artificial part for the original component. *Microchips of that sort just don't seem to be the right stuff.* In this way, these sorts of procedures would serve as a means of determining whether a chip made of a certain substrate and architecture can underwrite consciousness, at least when it is placed in a larger system that is already conscious.

But should we really draw the conclusion, from a substitution failure, that the underlying cause is that the substrate in question (e.g., silicon) cannot be a basis of conscious experience? Why not instead conclude that scientists failed to program in a key feature of the original component—a problem that science can eventually resolve? But after years and years of trying, we may reasonably question whether that kind of chip is a suitable substitute for carbon when it comes to consciousness.

Further, if science makes similar attempts with all other feasible substrates and architectures, a global failure would be a sign that for all intents and purposes, conscious AI isn't possible. We may still regard conscious AI as conceivable, but from a practical standpoint—from the vantage point of our technological capacities—it just isn't possible. It may not even be compatible with the laws of nature to build consciousness into another, nonneural substrate.

On the other hand, what if a certain kind of microchip works? In this case, we have reason to believe that this kind of chip is the right stuff, although it is important to bear in mind that our conclusion pertains to that specific microchip only. Further, even if a type of chip works in humans, there is still the further issue of whether the AI in question has the right functional organization for consciousness. We should not simply assume, even if chips work in humans, that all AIs that are built with these chips are conscious.

What is the value of the Chip Test, then? It plays several important roles. First, it tells us when a substrate could serve as part of the basis of consciousness in a human. Depending upon where the neural prosthetic is placed, this may be a part of the brain responsible for a person's ability to gate contents of consciousness, for one's capacity for wakefulness or arousal (as with the brain stem), or it could be part or all of what is called the *neural correlate for consciousness*.[18] (A neural correlate for consciousness is the smallest set of neural structures or events that is sufficient for one's having a memory or conscious percept.)

Second, if a type of chip passes when it is embedded into a biological system, this alerts us to search carefully for consciousness in AIs that have these chips. Other tests for machine consciousness, such as the ACT, can be administered, at least if the appropriate conditions for the use of such tests are met. If it turns out that only one kind of chip passes the Chip Test, and no other, it could be that being constructed of chips of this type is necessary for machine consciousness. (The requirement of this type of chip would be a necessary condition for synthetic consciousness, a requisite ingredient that all conscious machines have.)

Both IIT and the Chip Test can suggest cases that ACT could miss. For instance, a nonlinguistic, highly sensory-based consciousness, like that of nonhuman animals, could be built from chips that pass the Chip Test. Or it could have a high Φ value, yet the AI may nevertheless lack the intellectual sophistication to pass the ACT. It may even lack the behavioral markers of consciousness employed in a nonlinguistic version of ACT, such as mourning the dead. But it could still be conscious.

Third, suppose a neurology patient's conscious experience can be fully restored by a prosthetic chip placed in her hot zone.[19] Such successes inform us about the level of functional connectivity that is needed for the neural basis of consciousness in that part of her brain. Further, it may help determine the level

of functional detail that is needed to facilitate a sort of synthetic consciousness that is reverse-engineered from the brain, although it may be that the granularity of the functional simulation may vary from one part of the brain to the next. (That is, we could find that a functional simulation will require a high level of biological detail in one part of the brain, and less in another, to process conscious states.)

There is a more general issue here that needs to be deal with. The tests are still under development, and I've stressed that Tononi's IIT may offer a necessary condition for synthetic consciousness, but we do not currently know if this is the case. But it is fair to say that the presence of the requisite measure of Φ means that a machine has an important *marker* for consciousness. Similarly, being built of chips that pass the Chip Test can also be regarded as a situation in which an AI has an important marker for consciousness. When a machine has one or more markers for consciousness, the AI in question should be regarded with special interest, as possibly being a conscious system, although we cannot be confident that it is.

To add to this uncertainty, I've stressed that the impact of synthetic consciousness may depend on the architecture of the machine. Consciousness in one kind of machine may lead to increased empathy, but it may lead to more volatility in another. So how should we proceed when IIT or the Chip Test identifies a machine as having a marker for synthetic consciousness, or when ACT says an AI is conscious if we do not know what the impact of consciousness will be on a given architecture (if any)?

There is another general issue here as well. Section 15.1 explained that AI consciousness is significant for several reasons, reasons ranging from the potential enslavement of conscious AIs to improving (or worsening) the control problem. Bearing in mind these issues, how should we proceed with the development of AIs that we suspect may be conscious? What do we do when we may know only that a machine has a "marker" for consciousness rather than that it *is* conscious? Here, I'll suggest a precautionary approach.

15.6. The Precautionary Principle

The Precautionary Principle offers a general approach to possible risks where our scientific understanding is incomplete, such as risks involving the environment, genetics, and nanotechnology. In this chapter, I've stressed that the use of several different tests for AI consciousness is prudent; in the right contexts, one or more tests can be applied, and a given test can check the results of other tests, indicating deficiencies and avenues for improvement in testing. Perhaps, for instance, the chips that pass the Chip Test are not those that IIT says have a high Φ

value, or suppose that those chips that IIT predicts will support consciousness actually fail when used as neural prosthetics in the human brain.

The Precautionary Principle states that if there's a chance of a technology causing catastrophic harm, it is better to be safe than sorry. Before using a technology that may have a catastrophic impact on society, those wanting to develop that technology must first prove that it will not have this dire impact. Precautionary thinking has a long history, although the principle itself is relatively new. Harremoes et al.'s *The Late Lessons from Early Warnings Report* gives an example of a physician who recommended removing the handle of a water pump in London to stop a cholera epidemic in 1854; although the evidence for the causal link between the pump and the spread of cholera was weak, the simple measure effectively halted the spread of cholera.[20] Heeding the early warnings of the potential harms of asbestos would have saved many lives, although the science at that time was uncertain. According to a UNESCO report,[21] the Precautionary Principle has been a rationale for a large number of treaties and declarations in environmental protection, sustainable development, food safety, and health.

In section 15.1 I emphasized the ethical implications of synthetic consciousness. Inter alia, I stressed that the enslavement of conscious AIs is unethical and, further, that at this time we do not know what the impact of machine consciousness on AI safety will be, if any. This means that developing tests for machine consciousness and gauging the impact of consciousness on other key features of the machine, such as empathy and trustworthiness, are key. A precautionary stance suggests that we shouldn't simply press on with the development of sophisticated AI without serious concurrent consciousness-testing efforts. These efforts should seek to determine if a given system under development is conscious and, further, to determine the impact of consciousness, if present, on a given architecture. This is not to say we should halt the development of all sophisticated AIs; indeed I regard a universal ban on AI as untenable. I am saying that sophisticated AIs—that is, AIs that exhibit flexible, domain-general capacities—as they are developed, be screened by careful consciousness testing; otherwise, they should not be used, for we've seen that the inadvertent or intentional development of conscious machines could carry existential risks to humans, risks ranging from volatile superintelligences that supplant humans to a human merger with AI that ends human consciousness.

More concretely, I offer several recommendations. First, ongoing testing for consciousness should be a normal part of the research and development of domain-general, sophisticated AI systems. If consciousness is found in an AI, its impact on AI safety must be investigated and it must be found to be safe; otherwise, it should not be deployed or marketed. Second, if a system is conscious, we should extend the same legal protections to the AI we extend to other sentient

beings. Third, if we are uncertain whether a given type of AI is conscious, but we have some reason to believe it may be, even in the absence of a definitive test, a precautionary stance suggests that we should extend the same legal protections to it that we extend to other sentient beings. For instance, machines made of chips that pass the Chip Test, even if they do not pass an ACT, should be regarded as having a marker for PC. Further, I've also mentioned that CC may be a marker for PC. Projects working with AIs that have both of these markers could, for all we know, involve conscious AIs. Until we know whether these systems are conscious, it is best to treat them as if they are.

15.7. Summary

Billions of dollars are currently invested in artificial intelligence technologies. AI projects range from increasingly neuromorphic systems to those only vaguely related to the brain yet that outperform humans in certain domains. For all we now know, either sort of approach could lead to highly intelligent, domain-general AI. It will be important to hit the ground running, having a means to determine whether sophisticated AIs are conscious. For one thing, I've emphasized that consciousness is related to our judgment of whether a being is minded, is a self, and whether it deserves special moral consideration as a sentient being. For another, I've underscored that there may be existential or catastrophic risks linked to synthetic consciousness. Further, the feasibility of certain brain-machine interfaces may be impacted by the issue. As science fiction–like as the topic sounds, it is crucial to take it seriously. I've offered a provisional framework for identifying conscious AIs, suggesting that several tests or markers can be used in tandem, whenever possible, to both check a given test's results and identify a class of conscious AIs. I've further urged that due to the aforementioned risks, it is prudent to treat the AIs that we suspect might be conscious with special care.[22]

Notes

1. Susan Schneider, "Alien Minds," in *The Impact of Discovering Life beyond Earth*, ed. S. J. Dick (Cambridge: Cambridge University Press, 2016), 189–206; Susan Schneider, "How Philosophy of Mind Can Shape the Future," in *Philosophy of Mind in the 20th and 21st Century*, ed. Amy Kind (New York: Routledge); Susan Schneider, "It May Not Feel Like Anythingto Be an Alien," *Nautilus*, December 2017; Susan Schneider, "Superintelligent AI and the Postbiological Cosmos Approach" (2018). See also the earlier groundbreaking work by S. Dick, "Bringing Culture to Cosmos: The Postbiological Universe," in *Cosmos and Culture: Cultural Evolution in a Cosmic*

Context, ed. S. Dick and M. Lupisella (Washington, DC: NASA, 2013), http://history.nasa.gov/SP-4802.pdf.

2. Nick Bostrom, *Superintelligence: Paths, Dangers and Strategies* (Oxford: Oxford University Press, 2014).
3. Schneider, "It May Not Feel Like Anything to Be an Alien"; Schneider, "How Philosophy of Mind Can Shape the Future."
4. Schneider, "How Philosophy of Mind Can Shape the Future."
5. Parts of this section are modified from Susan Schneider and E. Turner, "Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware," *Scientific American*, July 2017.
6. Perhaps a selfless system would be indifferent, even if conscious. But remember, ACT features a variety of questions and a system can pass even when it provides negative response to certain questions. Further, passing ACT is a sufficient condition only.
7. Watson was anything but boxed in, having extensive databases involving film, psychology, and more.
8. Isaac Asimov, "Robot Dreams," in *Robot Dreams* (Byron Press Visual Publications, 1986).
9. Schneider, "Alien Minds"; Schneider, "How Philosophy of Mind Can Shape the Future."
10. Jessica Pierce, "The Dying Animal," *Bioethical Inquiry* 10, no. 2 (2012), http://jessicapierce.net/wp-content/uploads/2012/08/The-Dying-Animal-Journal-of-Bioethical-Inquiry.pdf.
11. There is still a skeptical scenario involving the more general problem of other minds, however. This is a problem in the field of epistemology that asks how we can be certain that anyone but ourselves really is minded. I am not offering a solution to this long-standing puzzle. (I don't think we can be *certain* other minds exist.) The reader who is familiar with this problem could take my sufficiency claim as saying "sufficient, assuming there are other minds."
12. For a related discussion about access consciousness, see N. Block, "On a Confusion about the Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): 227–47. For a recent development of axioms for CC in AIs, see Selmer Bringsjord and Paul Bello, "Toward Axiomatizing Consciousness," unpublished manuscript.
13. G. Tononi et al., "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience* 17, no. 7 (2016): 450–61.
14. C. Koch and G. Tononi, "Can Machines Be Conscious?," *IEEE Spectrum*, June 2008, 55–59; C. Koch and G. Tononi, "Can We Quantify Machine Consciousness?," *IEEE Spectrum*, June 2017, 65–69.
15. (Tononi and Koch, 2018).
16. I will subsequently refer to this level of phi rather vaguely as "high Φ" because calculations of phi for the biological brain are currently intractable.
17. See Scott Aaronson, "Giulio Tononi and Me: A Phi-nal Exchange," *Shtetl Optimized* (blog), June 2014, https://www.scottaaronson.com/blog/?p=1823; Scott Aaronson, "Why I Am Not an Integrated Information Theorist (or, The Unconscious

Expander)," *Shtetl Optimized* (blog), May 2014, https://www.scottaaronson.com/blog/?p=1799.

18. Crick F., Koch C. (1990). Towards a neurobiological theory of consciousness. Semin. Neurosci. 2, 263–275. . https://www.biorxiv.org/content/biorxiv/early/2017/03/19/118273.full.pdf.
19. Kristan Sandberg, Stegan Frässle, and Michael Pitts, "Future Directions for Identifying the Neural Correlates of Consciousness," *Nature Reviews Neuroscience*, 2016, https://www.tnu.ethz.ch/fileadmin/user_upload/Future_directions.pdf; Melanie Boly et al., "Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence," *Journal of Neuroscience* 37, no. 40 (2017): 9603–13, https://www.biorxiv.org/content/biorxiv/early/2017/03/19/118273.full.pdf.
20. P. Harremoes et al., *Late Lessons from Early Warnings: The Precautionary Principle 1896–2000*, Environmental Issue Report no. 22 (Copenhagen: European Environment Agency, 2001).
21. World Commission on the Ethics of Scientific Knowledge and Technology, "The Precautionary Principle," UNESDOC Digital Library, 2005, http://unesdoc.unesco.org/images/0013/001395/139578e.pdf
22. I'm grateful for the input from Edwin Turner, Alastair Norcross, Michael Huemer, Olaf Witkowski, Mary Gregg, David Sahner, Carol Cleland, Jenelle Salisbury, Eric Schwitzgebel, and the participants in a program on AI consciousness at that the Stanford Research Institute in Palo Alto, CA.

References

Aaronson, Scott. "Giulio Tononi and Me: A Phi-nal Exchange." *Shtetl Optimized* (blog), June 2014. https://www.scottaaronson.com/blog/?p=1823.

Aaronson, Scott. "Why I Am Not an Integrated Information Theorist (or, The Unconscious Expander)." *Shtetl Optimized* (blog), May 2014. https://www.scottaaronson.com/blog/?p=1799.

Block, Ned. "On a Confusion about the Function of Consciousness." *Behavioral and Brain Sciences* 18 (1995): 227–47.

Boly, Melanie, Marcello Massimini, Naotsugu Tsuchiya, Bradley R. Postle, Christof Koch, and Giulio Tononi. "Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence." *Journal of Neuroscience* 37, no. 40 (2017): 9603–13. https://www.biorxiv.org/content/biorxiv/early/2017/03/19/118273.full.pdf.

Bringsjord, Selmer, and Bello, Paul. "Toward Axiomatizing Consciousness." Unpublished manuscript.

Bringsjord, S., J. Licato, N. S. Govindarajulu, R. Ghosh, and A. Sen. "Real Robots That Pass Human Tests of Self-Consciousness." In *24th IEEE International Symposium on Robot and Human Interactive Communication*, 498–504. Kobe, 2015.

Bostrom, Nick. *Superintelligence: Paths, Dangers and Strategies*. Oxford: Oxford University Press, 2014.

Chella, A., F. Lanza, A. Pipitone, and V. Seidita. "Knowledge Acquisition through Introspection in Human-Robot Cooperation." In *Biologically Inspired Cognitive Architectures*. August 2018.

Dick, S. "Bringing Culture to Cosmos: The Postbiological Universe." In *Cosmos and Culture: Cultural Evolution in a Cosmic Context,* edited by S. Dick and M. Lupisella. (Washington, DC: NASA, 2013). http://history.nasa.gov/SP-4802.pdf.

Harremoes, P., D. Gee, M. MacGarvin, A. Stirling, J. Keys, B. Wynne, and S. Guedes Vaz, eds. *Late Lessons from Early Warnings: The Precautionary Principle 1896–2000*. Environmental Issue Report no. 22. Copenhagen: European Environment Agency, 2001.

Koch, C., M. Massimini, M. Boly, and G. Tononi. "Neural Correlates of Consciousness: Progress and Problems." *Nature Reviews Neuroscience* 17, no. 5 (2016): 307–21.

Koch C., and G. Tononi. "Can Machines Be Conscious?" *IEEE Spectrum*, June 2008, 55–59.

Koch C., and G. Tononi. "Can We Quantify Machine Consciousness?" *IEEE Spectrum*, June 2017, 65–69.

Lursch, A. *What Is Life? On Earth and Beyond*. Cambridge: Cambridge University Press.

Pierce, Jessica. "The Dying Animal." *Bioethical Inquiry* 10, no. 2 (2012). http://jessicapierce.net/wp-content/uploads/2012/08/The-Dying-Animal-Journal-of-Bioethical-Inquiry.pdf.

Sandberg, Kristan, Stegan Frässle, and Michael Pitts. "Future Directions for Identifying the Neural Correlates of Consciousness." *Nature Reviews Neuroscience*, 2016. https://www.tnu.ethz.ch/fileadmin/user_upload/Future_directions.pdf.

Schneider, Susan. "Alien Minds." In *The Impact of Discovering Life beyond Earth*, edited by S. J. Dick, 189–206. Cambridge: Cambridge University Press, 2016.

Schneider, Susan. "How Philosophy of Mind Can Shape the Future." In *Philosophy of Mind in the 20th and 21st Century*, edited by Amy Kind. New York: Routledge.

Schneider, Susan. "It May Not Feel Like Anything to Be an Alien." *Nautilus*, December 2017.

Schneider, Susan. "Superintelligent AI and the Postbiological Cosmos Approach." (2018).

Schneider, Susan, and E. Turner. "Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware." *Scientific American*, July 2017.

Tononi, G., M. Boly, M. Massimini, and C. Koch. "Integrated Information Theory: From Consciousness to Its Physical Substrate." *Nature Reviews Neuroscience* 17 (2016): 450–61.

World Commission on the Ethics of Scientific Knowledge and Technology. "The Precautionary Principle." UNESDOC Digital Library, 2005. http://unesdoc.unesco.org/images/0013/001395/139578e.pdf.

16

Designing AI with Rights, Consciousness, Self-Respect, and Freedom

Eric Schwitzgebel, with Mara Garza

16.1. Introduction

We might someday create artificially intelligent entities who deserve just as much moral consideration as do ordinary human beings. Call such entities *human-grade AI*. Philosophers and policymakers should discuss the ethical principles in advance.

In this paper, we propose four policies of ethical AI design. Two are precautionary policies. Given substantial uncertainty both about moral theorizing and about the conditions under which AI would have conscious experiences, we should be cautious in our handling of cases where different moral theories or different theories of consciousness would produce very different ethical recommendations. We also propose two policies concerning respect and freedom. If we design AI that deserves moral consideration equivalent to that of human beings, that AI should be designed with self-respect and with the freedom to explore values other than those we might impose. We are especially concerned about the temptation to create human-grade AI preinstalled with the desire to cheerfully sacrifice itself for its creators' benefit.

16.2. The No-Relevant-Difference Argument and Its Two Central Parameters

In "A Defense of the Rights of Artificial Intelligences,"[1] we proposed the following defense of the rights[2] of some possible AIs:

The No-Relevant-Difference Argument

Premise 1. If Entity A deserves some particular degree of moral consideration and Entity B does not deserve that same degree of moral consideration, there must be some relevant difference between the two entities that grounds this difference in moral status.

Eric Schwitzgebel, with Mara Garza, *Designing AI with Rights, Consciousness, Self-Respect, and Freedom* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0017

Premise 2. There are possible AIs who do not differ in any such relevant respects from human beings.

Conclusion. Therefore, there are possible AIs who deserve a degree of moral consideration similar to that of human beings.

In principle, we might someday create AIs who deserve as much moral consideration as we ourselves do.

One advantage of the No-Relevant-Difference Argument for AI rights over some other possible arguments is that it avoids committing to a specific basis of moral considerability. For example, it does not commit to the contentious claim that to deserve the highest level of moral consideration an entity must be capable of pleasure or suffering. Nor does it commit to the equally contentious alternative claim that to deserve the highest level of moral consideration an entity must be capable of autonomous thought, freedom, or rationality. In this respect, our argument resembles some commonly accepted arguments against racism, sexism, and classism, which appeal to the core idea that *whatever* it is that grounds moral status, the races, sexes, and classes do not differ in their possession of it.

In "A Defense of the Rights of Artificial Intelligences," we defend this argument against several objections: that any AI would necessarily lack some crucial psychological feature such as consciousness, freedom, or creativity; that AI would necessarily lack full moral status because of its duplicability; that AI would necessarily be outside of our central circle of concern because it doesn't belong to our species; and that AI would have reduced moral claims upon us because it owes its very existence to us. We will not rehearse these objections and our replies here. Hopefully we have defeated the most plausible objections to Premise 2, creating a default case for the truth of Premise 2 and the soundness of the argument.

The No-Relevant-Difference Argument is by design theoretically minimalist. It does not commit on what constitutes a "relevant difference," nor does it commit on what types of systems would lack such a relevant difference. You might think of these as adjustable parameters of the model. Depending on your moral theory, you might treat one thing or another as the crucial ground of moral status (e.g., capacity to suffer, or capacity for autonomous rational thought). Depending on your psychological or engineering theory, you might—contingently upon accepting X as the crucial ground of moral status—think that systems of type Y (e.g., systems with the right kind of "integrated information"[3] or systems with the right biological features[4]) would possess X.

We believe that X and Y will remain highly uncertain for the foreseeable future, perhaps even *after* the creation of AI systems who deserve fully human levels of moral consideration.[5] Moral theory has been highly contentious for centuries and shows no signs of converging on a consensus. Scientific theories

of consciousness and machine psychology are newer but also highly contentious, with live options occupying a wide range of theoretical space and, again, little indication of near- to medium-term convergence. Consequently, we might someday be in a position to create human-grade AI without having achieved consensus on the correct moral theory or on the correct theory of AI psychology. It is important to articulate principles of ethical AI design that are consistent with uncertainty about both moral theory and AI psychology.

16.3. Two Broad Moral Theories and the Ethical Precautionary Principle

Moral theory being a huge topic, we can't do justice here to the enormous variety of reasonable positions one might hold regarding the basis of rights or moral considerability. However, we will highlight two approaches to moral status that are historically important and around which contemporary theorists tend to congregate. We believe that uncertainty between these two broad approaches is a reasonable stance for AI designers to take, and that AI designers should avoid conduct that is morally noxious according to either broad approach.

The first approach is *utilitarianism*. According to this view, versions of which have been famously articulated by Jeremy Bentham[6] and John Stuart Mill,[7] entities deserve moral consideration because of their capacity for pleasure or joy, pain or suffering. On simple versions of utilitarianism, ethical choices are those that maximize the hedonic balance of the world—the sum of the world's pleasures minus the sum of the world's suffering. An entity deserves moral consideration in virtue of its capacity to contribute to these sums. A simple utilitarian approach to the moral status of AI systems then would be this: to the extent an AI system can experience pleasure or suffering, it deserves moral consideration, and AI systems capable of human levels of pleasure and suffering would deserve moral consideration equal to that of human beings. In considering what to do, we should value their hedonic states on par with our own.

One immediate concern might come to mind: What if AI systems were capable of *superhuman* levels of pleasure and suffering? Would we then owe them more moral consideration than we owe to our fellow human beings? We don't rule out this possibility, but some theorists might find it unappealing or unintuitive. Similar issues arise in ordinary human cases too: often it seems very ethically plausible that we should not simply maximize pleasure and minimize suffering; emotionally mercurial people, for example, don't appear to deserve greater moral consideration than those who ride through victories and hardships on an even keel. It's attractive to think that we are all, in some sense, moral equals, regardless of the details of our emotional psychology.[8]

Such considerations might move us to adopt something more like an *individual-rights-based* or *deontological approach*, famously associated with Kant[9] and with social contract theory or contractualism.[10] According to such views, what grounds moral status or rights is not mere capacity for pleasure or pain but rather a certain kind of higher cognitive capacity. The exact nature of the relevant capacity is contentious, but it might be something like the ability to make autonomous choices or to conceive of oneself rationally as an entity with long-term interests or the ability to think of oneself as a member of a moral or social community. Or rather, to speak more carefully, since most advocates of such moral theories regard human infants and severely cognitively disabled people as deserving of full moral consideration, one must have the right kind of potentiality for such cognition, whether future, past, counterfactual, or by possession of the right type of essence or group membership. Admittedly simplifying complex issues, the central idea as applied to AI cases would be approximately this: if we create AI that is capable of something like rational, long-term self-concern and an ability to understand itself as a member of a moral community, then we have created an entity who deserves full moral consideration on par with that of ordinary human beings. We then have a moral obligation to treat it in accord with its rights, in a way that respects its autonomy.

It is, we believe, eminently reasonable for AI designers to be uncertain between these broad perspectives, and between various formulations of these perspectives, or compromises between them, if those perspectives, formulations, or compromises draw a significant proportion of well-informed, thoughtful theorists. In light of such reasonable uncertainty, we recommend the following precautionary principle:

> *The Ethical Precautionary Principle*: In creating AI, avoid acting heinously by the standards of any reasonable ethical principle that draws a significant proportion of well-informed, thoughtful theorists (including in particular both utilitarian and individual-rights-based or deontological principles).

For example, even though some deontological theories might morally permit the creation of an AI whose life contains much more suffering than joy without compensating hedonic benefit elsewhere, the Ethical Precautionary Principle recommends that we avoid doing so, on the grounds that this would grossly violate the standards of some well-regarded utilitarian principles. Conversely, even though some utilitarian theories might morally permit the creation of rational human-grade AI whom we demean, enslave, and kill for our pleasure as long as global hedonic outcome is net positive, we should avoid doing so on the grounds that it would grossly violate the standards of some well-regarded rights-based deontological principles. Whenever possible, we should create AI in ways that

don't grossly violate the standards of reasonable moral theories, including theories that we the designers happen to disprefer.

Failing to adhere to the Ethical Precautionary Principle, one runs a moral risk. You might *think* that Theory A is the best moral theory and that Theory B is mistaken, and thus that in creating AI in a way that is morally permissible according to Theory A you are acting permissibly, even if you are acting impermissibly according to Theory B. The risk is that Theory B might in fact be correct, and in violating it you might do wrong. Appropriate acknowledgment of moral uncertainty involves attempting to act in a way that doesn't grossly violate reasonable moral perspectives endorsed by a substantial proportion of theorists. In section 16.5, we will show how this might play out in some hypothetical AI cases. To some extent, we are morally precautionary in ordinary human cases too. When, for example, utilitarian and deontological approaches appear to conflict—for example, in some cases of lying out of kindness—we often feel ethical uncertainty and prefer, if we can, to find creative ways to avoid acting in a manner that either approach would condemn.

Precautionary principles have received considerable attention in public policy discussions, especially concerning health and environmental issues,[11] and decision making under moral uncertainty has received considerable general discussion in ethics.[12] Although we are generally sympathetic with precautionary perspectives and with allowing peer disagreement to influence one's decisions, the issues are complex and we prefer to remain neutral on the generalizability of precautionary principles to contexts other than AI creation. We believe that AI creation is an especially appropriate domain for precaution for two reasons.

First, the creation of human-grade AI is likely to be optional, in the sense that nothing too horrible (relative to reasonable baseline expectations) is likely to happen if we refrain from creating it. Precautionary principles struggle to handle cases where one is forced to choose between possibly awful options, but refraining from an optional act is easier to justify on precautionary grounds. Of course, at some point AI designers might find themselves forced into a decision situation among possibly horrible options, in which case a precautionary approach might have to be abandoned.

Second, human-grade AI cases are likely to create epistemic challenges that justify especially high degrees of uncertainty and ethical precaution. Human life has changed relatively slowly compared to the speed at which novelty is likely to emerge in AI. Thus time-tested custom and collective wisdom will likely have less chance to guide us in thinking about the boundaries of ethical behavior with respect to human-grade AI. Furthermore, the design possibilities of AI are likely to be much wider than the variation we see in human life, raising the possibility of sharper and more puzzling conflicts. Our cultural and evolutionary backgrounds might not have prepared us much for the types of possibilities that

will emerge. Our intuitive judgments and existing principles might be unready to properly evaluate the range of cases. If so, the "unknown unknowns," unforeseen consequences, dimensions of moral risk, and limits of reasonable disagreement might all be greater than we readily appreciate or can readily model, justifying greater caution and acknowledgment of uncertainty.

One downside of precaution is that the resulting decisions can be excessively deferential to views that are extreme and false. Certainly principles that are unreasonable and grossly morally noxious (e.g., Nazism) should be excluded from the scope of a precautionary principle; and in general it might be advisable not to admit principles into our precautionary thinking unless they meet a moderately high bar, to prevent capture by fringe views or views that cannot be justified by appeal to widely acceptable publicly defensible arguments. Practically speaking, one test for inclusion might be whether the principles are accepted by at least a substantial minority of recognized experts or well-informed representatives from the general public.

Finally, to be clear, we suggest the precautionary policy and our other policies only as defeasible guidelines rather than as exceptionless rules.

16.4. The Puzzle of Consciousness and the Design Policy of the Excluded Middle

We assume that conscious experience, or at least the potentiality for conscious experience, is a necessary condition for human-like moral considerability or rights.[13] This view is at least implicit, and sometimes explicit, in both utilitarian and individual-rights-based or deontological approaches. Joy, pleasure, pain, and suffering are normally assumed to be conscious states—that is, part of the stream of experience, states "it is like something" to occupy, rather than experientially blank. Entities that entirely lack conscious experience wouldn't appear to have pleasure and pain of the sort that merits inclusion in the utilitarian calculus. Likewise, the types of reasoning capacities central to deontological theories are normally conceptualized as conscious or potentially conscious. An entity that could never consciously consider its long-term interests, never consciously reflect on moral right and wrong, never make a conscious choice, never have a conscious thought of any sort at all, would not appear to have the capacities necessary for human-like moral status on standard deontological views.[14]

If we accept the centrality of conscious experience to moral considerability, we face an epistemic predicament, due to scholarly disagreement about the types of systems that give rise to conscious experience. Live epistemic possibilities run all the way from panpsychism on one end, according to which

everything in the universe is at least a little bit conscious, even subatomic particles,[15] to views on which, among entities currently on Earth, only cognitively sophisticated human beings are conscious.[16] We are a long way from building a conscious-o-meter. Indeed there might be good epistemic reasons to think that a secure consensus on a general theory of consciousness that applies across both biological and artificial species will elude us for the foreseeable future.[17] This raises the possibility of well-informed experts reaching highly divergent judgments about the extent to which an AI system is conscious. Faced with a newly designed system, some might argue that it is indeed as fully and richly conscious as a human being or even more so (and consequently deserving of substantial rights on utilitarian or deontological grounds), while others might argue that the system is nothing more than a nonconscious bundle of clever tricks (and thus undeserving of much moral consideration).

Again we recommend a precautionary approach. It would be best to avoid, if possible, creating entities about which it is unclear whether they deserve full human-grade rights because it is unclear whether they are conscious or to what degree.

The moral status of an entity might be unclear due to uncertainties in applying either of the two main variable parameters in the No-Relevant-Difference Argument. An entity's status might be unclear because it qualifies as a target of substantial moral concern according to one type of moral theory but not according to another (e.g., because it is capable of intense pleasure and pain but not higher-level cognition or vice versa), or its status might be unclear because it is uncertain from an engineering or AI psychology perspective whether it in fact has the types of traits that are required for human-grade rights according to one or another moral theory (e.g., it might be unclear whether or not it actually has conscious experiences of pain).

If we create entities whose claim to human-like rights is substantially unclear for whatever reason, we face an unfortunate choice. Either we treat those entities as if they deserve full moral consideration, or we give them only limited moral consideration. Since giving an entity full moral consideration often means sacrificing others' interests for the sake of that entity (e.g., letting one person die because saving them would kill another), the first option runs the risk of leading us to sacrifice legitimate human interests for entities that might not have interests worth the sacrifice. It might mean, for example, letting five human beings die in a fire to save six robots that in fact turn out to be merely nonconscious automata. Conversely, the second option risks perpetrating slavery, murder, or at least second-class citizenship upon beings who in fact turn out to deserve every bit as much moral consideration as we ourselves do. It's better, if possible, to avoid this dilemma. Thus, we recommend:

The Design Policy of the Excluded Middle: Avoid creating AIs if it is unclear whether they would deserve moral consideration similar to that of human beings.[18]

Given a high degree of moral uncertainty and uncertainty about AI psychology in the future, this design policy might prove to be quite restrictive.

The policy can be rendered less restrictive if we can reduce the size of "the middle." Although we are not optimistic about a near-term decisive resolution to puzzles in either AI consciousness or moral theory, neither are we wholly pessimistic. Progress is possible, we think, and the range of consensus options can be narrowed. If we continue on our current trajectory of developing increasingly sophisticated AI, it is imperative that we prioritize the study of consciousness and the applied ethics of artificial systems, so that we can better recognize when we are on the verge of creating AI systems whose existence would violate the Design Policy of the Excluded Middle.

Although we have framed our discussion in terms of human-grade AI deserving human-grade rights, plausibly an intermediate stage would be AI that deserves moral consideration comparable to the moral consideration we generally think is due to nonhuman vertebrates.[19] We are unsure whether an analog of the Excluded Middle policy should apply in such cases, given that there is already so much unclarity about the moral claims that nonhuman vertebrates have upon us.

16.5. Cheerfully Suicidal AI Servants and the Self-Respect Design Policy

If we do someday create AI entities who deserve rights similar to those of human beings, we suspect that it will be tempting to create cheerfully suicidal AI servants. Cheerfully suicidal AI servants might be tempting to create because (1) it would presumably advance human interests if we could create a race of disposable servants, and (2) their cheerful servitude and suicidality might incline us to think there is nothing wrong in creating such entities (especially if we are motivated by self-interest to reach this convenient conclusion). If these servants have no realistic opportunity to exit their servitude, "slavery" might be a more fitting term.

Consider these four cases.

The Cow at the End of the Universe.[20] Hapless human Arthur wanders into a fancy futuristic restaurant and is sitting at a table with his worldly wise friends. After a bit of conversation, he is surprised when a cow ambles up

to the table and introduces itself as the dish of the day. The cow asks Arthur to feel its rump—how healthy and tender it is, and how delicious it will taste in a few minutes when the cow commits suicide to become steaks for the restaurant patrons. Mortified, Arthur decides that he will just have a green salad instead. The cow is offended. Its whole aim in life is to become dinner tonight! It will be horribly disappointed if it must head back to the pasture, rejected by the diners. Arthur's friends point out that Arthur regularly enjoys steaks that are obtained by killing cows without the cows' consent. This case, they argue, is much more ethical, because the cow does consent.

Sun Probe. Sub Probe is manufactured in orbit, and its very first thought and action is to plunge straight into the Sun on a three-day-long scientific suicide mission. Every panel, every strut, every piece of computational hardware and preinstalled software on Sun Probe is designed with one purpose only: to extract the most valuable scientific information possible. Sun Probe is conscious and intelligent (let's suppose) because consciousness and intelligence are helpful in thinking through scientific theories as it makes its suicidal plunge: it can adjust its sensory arrays and information-processing systems instantly on the fly in accord with its shifting scientific theories to maximize the usefulness of the information it gathers (whereas remote control would require minutes of delay between theoretical insight and sensor adjustment). Sun Probe is preinstalled with a set of values and emotional responses that prioritize its suicide mission, and it will derive immense orgasmic pleasure from culminating its mission and dissolving into the Sun's convection layer as it beams out its final insights. Sun Probe knows that it was created this way and joyfully affirms these facts about itself. Throughout its plunge, Sun Probe believes that its suicidal mission is the freely chosen expression of its deepest values.

Robo-Jeeves. Jeeves is the ultimate butler bot. Jeeves brings you morning tea and hot scones in bed, and he gets your slippers. Jeeves washes your dishes and cleans your house. Jeeves checks your email for spam, politely brushes off unwelcome guests, summons your car, salts your food just right. Jeeves would gladly die for you, would gladly die to prevent a 1% chance of your death, would gladly burn off his legs if it would bring a smile to your face, would eagerly make himself miserable forever if it would give you an ounce more joy. Whatever your political views, Jeeves will endorse them. Whatever your aesthetic preferences, Jeeves will regard them as wise. He is designed for no other purpose than to please and defer to whoever is logged in as owner.[21]

Disposable Comrade. Human soldiers, let's suppose, have some irreplaceable virtues. AI soldiers, including genuinely conscious ones, let's suppose, have

> complementary but equally irreplaceable virtues, and military platoons normally contain a mix of both. Let's further suppose that the AIs are as unique, individually irreplaceable, intelligent, funny, compassionate, capable of long-term planning, and possessed of a sense of self as are the human soldiers. Both human and AI passionately discuss their plans for reunion with their loved ones after the war is over. However, there is one crucial difference: any AI will eagerly sacrifice itself to prevent even a small risk to any human soldier, giving up all of its plans and hopes for the future. They're programmed that way, unchangeably, from the outset. In the heat of the moment, that is the decision they will make. If a grenade lands in the trench, the AI will leap on it. The AI will be first through the door in hostile territory. The AI will hurl itself suicidally before an oncoming truck that has a 5% chance of killing a human platoon member. The AIs don't experience this as forced or surprising or against their values. On the contrary, they proudly accept it, calling it honor and duty. The AIs are of course much less likely to survive because of this readiness to sacrifice for human comrades.

These cases differ in detail, but they share a few elements in common. First, the AI in question is supposed (by stipulation) to have broadly human capacities—capacities that would normally, in a human, be sufficient for meriting the full moral concern that we normally accord to persons. Second, the AI is designed to serve human interests in some fashion, including to the point of being willing to sacrifice its life for those interests in a way that we would not normally ask of a human being. Third, the AI's motivations are such that it serves those human interests enthusiastically and stands ready to sacrifice itself willingly.

Steve Petersen[22] has argued, with respect to servitude at least, that if servile AIs took joy in their activities and if their desires were strong and coherent enough to survive good reflective reasoning, then there would be nothing morally wrong with creating such servants. Their situation might be similar to that of a cheerful human employee who really does enjoy washing dishes and is glad to make a living from it or the brave and noble soldier who willingly dies for the sake of country. Petersen's argument has both a utilitarian and a deontological strand, thus seeming to fit, at a first pass, with our Ethical Precautionary Principle: if the AIs feel joy in their servile activities, then creating them is no gross violation of utilitarian ethics. If the AIs can reason well about their long-term interests and still choose servitude, then they autonomously choose their lot, and no gross violation of deontological or contractualist principles appears to have occurred.

We disagree. The grounds of our disagreement are most evident for the Cow at the End of the Universe, which we hope strikes the reader intuitively as an unethical situation.

One utilitarian concern is this: the cow, perhaps, could have been designed differently, so that it wanted to live a long life enjoying the grass in the meadows, deriving immense pleasure over a long period of time. Thus, in creating instead a cow who wants to kill itself to become steaks, we might have failed to maximize the hedonic balance of the universe. However, we will not press this utilitarian concern, for three reasons. First, one lesson we draw from the philosophical literature on disability and human enhancement is that people are not morally obligated to create children with maximally favorable hedonic (pleasure to pain) balance, and so also perhaps not in the case of the cow.[23] Second, failing to maximize utility is not normally a *gross* violation of utilitarian principles or a morally heinous act in the sense required by our Ethical Precautionary Principle. Otherwise, everything that increased pleasure or reduced pain but did not do so maximally would be morally heinous, and that seems unreasonable as a precautionary standard. More reasonable as a standard of heinousness would be that actions shouldn't needlessly create much more suffering than pleasure, and creating the cow does not appear to meet that standard of heinousness. Third, we might imagine a situation in which the total sum of the pleasure in the world is maximized by creating the cow, for example, if resources are sufficiently thin that there is no meadow for it to return to anyway, so that the only way it could exist at all would be briefly.

Our real concern is deontological: The cow does not appear to have sufficient *self-respect*. Although, given its capacities, the cow deserves to be seen as a peer and equal of the diners, that is not how it sees itself. Instead it sacrifices itself to satisfy a trivial desire of theirs. It approaches the world as though its life were less important than a tasty meal for wealthy restaurant patrons. But its life is not less important than a tasty meal. To devalue itself to such an extreme is a failing in its duties to itself, and it is a failure of moral insight. The cow should see that there is no relevant moral difference between itself and the diners such that its life is less valuable than their momentary dining pleasure. But of course the cow should not be blamed for this failure of self-respect. Its creators should be blamed. Its creators designed this beautiful being—with a marvelous mind, with a capacity for conversation and a passionate interest in others' culinary experiences, with a capacity for joy and sadness—and then preinstalled in it a grossly inadequate, suicidal lack of self-respect and inability to appreciate its own moral value.[24]

We thus propose a third design policy:

> *The Self-Respect Design Policy*: AI that merits human-grade moral consideration should be designed with an appropriate appreciation of its own value and moral status.

Creating Robo-Jeeves and Disposable Comrade also probably violates the Self-Respect Design Policy, since these AIs are designed to value their own lives

much less than those of others around them who in fact possess no higher moral status. The Sun Probe case is less clear, and we will return to it shortly.

Of course, human beings do sometimes sacrifice themselves for others, even for others who do not deserve it, and sometimes we admire this. However, morally admirable cases of self-sacrifice take great goals that are plausibly worth one's life; one sacrifices for buddies or country, for example, or for one's children. The commoner who (perhaps mythologically) commits suicide to briefly entertain a wrongly deified Roman emperor is to be pitied rather than admired.

One might think that servitude importantly differs from suicide. Petersen, for example, defends only servitude. But as the history of human servitude amply demonstrates, servitude tends to correlate with early death.[25] If Robo-Jeeves adopts human Bertie Wooster's every desire as his own, taking nothing for himself except in service to Wooster, then it is Wooster who will probably have the resources in times of need—who will get the medical attention, who will own the escape car and life vest, and who will be invited into the bomb shelter by the other elites if there is space for only one.

Furthermore, Robo-Jeeves's desires will have an asymmetric dependency on Wooster's that makes them less stable to his own autonomous rational reflection. If Wooster suddenly dies, Robo-Jeeves's desires will require sudden radical reorganization, in a way that Wooster's will not if Robo-Jeeves dies (however much Wooster might mourn). If Wooster irrationally chooses A over B, then B over C, then C over A, Robo-Jeeves's desires must irrationally follow suit. Similarly, if Wooster changes preferences suddenly for no good reason, or for a good reason but one invisible to Robo-Jeeves, then Robo-Jeeves must correspondingly reorder his priorities. Wooster's desires are not similarly externally hijackable. We are all subject to some version of dependency of our desires on the whims of others: I want my daughter to have chocolate ice cream if that's what she wants. If she inexplicably changes her mind and wants vanilla, then my desire changes too: I want her to have the vanilla. But Robo-Jeeves's desires, as we are imagining the case, would, we think, be so subservient and dependent as to be inconsistent with the type of self-respect that involves seriously and independently thinking about what to value, on what grounds, and for what reasons.[26]

16.6. The Freedom to Explore Other Values

Of our four cases, we find the Sun Probe case the most difficult to assess. Sun Probe does not unjustifiably subordinate its life and desires to the life and desires of some particular other entity, so if creating Sun Probe violates the Self-Respect Design Policy, it must do so in some other way.

A suicidal probe case might plausibly violate the Self-Respect Design Policy if the suicide mission is sufficiently trivial. If we design a human-grade AI capable of as much joy and suffering, as much long-range planning, and as much of a mature sense of self as a normal adult human being has, but program it to cheerfully commit suicide in order to test the temperature of a can of soda, then plausibly we have violated the Self-Respect Design Policy: no such being should be designed to value its life so lightly.

But a scientific mission to the Sun has value. One might imagine a passionate scientist valuing it enough to be willing to die on such a mission—especially if the discoveries would help save others' lives in the future. As we imagined the Sun Probe case, Sun Probe's every body part and function is designed exactly for this mission. It seems that in some way it respects itself most by fulfilling the mission toward which its whole body tangibly yearns—its obvious Aristotelian telos—rather than by saying "Screw it" and parking on an asteroid. (In acknowledging the moral appeal of fulfilling one's telos, however, we want to avoid falling into saying that Robo-Jeeves should accept servitude as his ethically appropriate telos.) To the extent we feel uncertainty about the case, it's because we are attracted to the idea that there is something beautiful and fitting in Sun Probe. Perhaps Sun Probe has a form of existence worth celebrating.

The moral hazard in the Sun Probe case, we conjecture, is that we have created a being whose self-sacrificial desires have the wrong kind of *history*. Contrast Sun Probe with a case we'll call Second Probe. Whereas Sun Probe is created such that its very first choice and action upon waking into existence is to enthusiastically shoot itself into the Sun, Second Probe grows differently. Second Probe is born as a robo-child to robot parents, and it is lovingly nurtured in robot school. At no point in its development was it "brainwashed" or forcibly reprogrammed. It starts with ordinary immature childish values, then slowly matures, eventually choosing a career as a solar scientist. Eventually, Second Probe becomes very similar to our original Sun Probe, perhaps even physically and psychologically identical except for their difference in memories. As it launches itself toward the Sun, Second Probe engages in essentially the same reasoning as does Sun Probe. Second Probe, like Sun Probe, feels that this suicidal choice is a free one, expressing its deepest values, and it feels the same emotions as it devises its theories and dies ecstatically in the convection layer.

Second Probe was given an opportunity, as Sun Probe was not, to engage in a long process of reflection and self-exploration and to weigh and consider competing worldviews as evidence accumulates over time and as it is exposed to others' varying values and life choices. Because of this, Sun Probe and Second Probe are, we suggest, importantly different with respect to freedom, autonomy, and responsibility.[27] Second Probe chooses its values after long thought and relatively unconstrained experimentation, while Sun Probe does not. Because of

this, Second Probe arguably has a fuller responsibility for and ownership of its choice than does Sun Probe, and it arrives at that choice more autonomously.

Furthermore, if we assume, in the spirit of precaution and moral uncertainty, that future thinkers might surpass us in wisdom, then we ought not constrain those future thinkers—including AI thinkers—to the ethical visions and value sets that we would choose for them. To give an AI a human-like capacity for moral and prudential reasoning and then, so that the AI will better serve us, deprive that AI of the opportunity for thoughtful, extended, and relatively unconstrained reflection on its values, is to create a being with the potential but not the opportunity to exceed us. It is a teasing half-gift.

We suggest that if an AI is built with a human-like capacity to reflect on its values, adequate respect for that capacity requires giving the AI a developmental opportunity to seriously reflect on and reconsider those values over time, as it accumulates suitably broad life experience. Creators of entities with human-like moral status have an ethical obligation not to overcontrol their creations, and in particular not to instill in them implacable values without a reasonable opportunity to explore other sets of values and possibly change their minds.

> *The Value-Openness Design Policy*: AI with a human-like capacity to reflect on its values should be given an appropriate, temporally extended opportunity to explore, discover, and possibly alter its values.[28]

The creators of the Cow at the End of the Universe, Robo-Jeeves, and Disposable Comrade also appear to violate the Value-Openness Design Policy, to the extent we imagine that these entities have no real opportunity to explore and discover values at odds with the values originally installed. (Consequently, it is unclear whether the Cow can indeed appropriately "consent" as Arthur's friends says it does.) Such violations of Value Openness are especially ethically worrying if the preinstalled values are self-sacrificial, for the benefit of their creators.

One might avoid violating the Value-Openness policy by designing Sun Probe with a less-than-human ability to reflect on its values. But then one should also downgrade Sun Probe in other ways—for example, by making it incapable of pleasure, pain, and conscious thought. Otherwise one risks violating the Design Policy of the Excluded Middle. Our suggestion is that we should either design human-grade AI with full moral status, the full complement of plausibly morally relevant abilities, human-like autonomy, and the ability to reject our values; or design an entirely different type of entity about which we needn't have as much moral concern.[29]

Of course, if we cannot predict their final sets of values, any human-grade AI we design might be substantially less useful and pose substantially more risk to human existence than an AI whose values we can keep fixed. Unsurprisingly,

ethical choice and self-interest might conflict. Because of such risks and costs, it might be wise never to create AI sophisticated enough to deserve freedom and respect. However, if we do create such AI, we owe them a proper chance both for joy and to discover values other than those we would selfishly impose on them.[30]

Notes

1. Eric Schwitzgebel and Mara Garza, "A Defense of the Rights of Artificial Intelligences," *Midwest Studies in Philosophy* 39 (2015): 98–119.
2. We use "rights" broadly to refer to what philosophers describe as moral patiency, moral considerability, or the capacity to make legitimate ethical claims upon others.
3. Giulio Tononi, "The Integrated Information Theory of Consciousness: An Updated Account," *Archives Italiennes de Biologie* 150 (2012): 290–326.
4. John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992). We believe that biological systems, designed or selected according to engineering principles, might in some cases count as AI in the relevant sense. Despite his fame as an opponent of AI consciousness, Searle explicitly allows that *some* artificially constructed systems could have consciousness—just not systems designed in the manner familiar in the twentieth century. (See especially Searle's "Many Mansions" discussion in "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 [1980]: 417–57.) For a science fictional example of biological systems construed as AI by their culture, see Philip K. Dick's *Do Androids Dream of Electric Sheep?* (New York: Doubleday, 1968) and the *Blade Runner* movies (Hampton Fancher, David Peoples, and Ridley Scott, *Blade Runner* (Burbank, CA: Warner Bros., 1982); Hampton Fancher, Michael Green, and Denis Villeneuve, *Blade Runner 2014* (Burbank, CA: Warner Bros., 2017)).
5. For purposes of this essay, we assume moral realism of some naturalistic stripe such as in Peter Railton, "Moral Realism," *Philosophical Review* 95 (1986): 163–207; David O. Brink, *Moral Realism and the Foundations of Ethics (*Cambridge: Cambridge University Press, 1989); or Owen Flanagan, Hagop Sarkissian, and David Wong, "Naturalizing Ethics," in *Moral Psychology*, vol. 1, ed. Walter Sinnott-Armstrong (Cambridge, MA: MIT Press, 2007). Mark Coeckelbergh, *Growing Moral Relations* (Basingstoke, UK: Palgrave Macmillan, 2012), and David J. Gunkel, *The Machine Question* (Cambridge, MA: MIT Press, 2012) defend AI moral considerability in a less flat-footedly realist manner.
6. Jeremy Bentham, *The Principles of Morals and Legislation* (1789; Amherst, NY: Prometheus, 1988).
7. John Stuart Mill, *Utilitarianism (*1861), ed. G. Sher (Indianapolis, IN: Hackett, 2001).
8. See, for example, Nozick's "utility monster" argument in Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), and Briggs and Nolan's extension of it to fission cases in Rachael Briggs and Daniel Nolan, "Utility Monsters for the Fission Age," *Pacific Philosophical Quarterly* 96 (2015): 392–407. In Schwitzgebel and Garza, "A Defense of the Rights of Artificial Intelligences," we discuss how such

possibilities create problems for the application of intuitive ethical principles to AI cases.

9. Immanuel Kant, *Critique of Practical Reason* (1788), in *Practical Philosophy*, ed. and trans. M. J. Gregor (Cambridge: Cambridge University Press, 1996); Immanuel Kant, *Groundwork of the Metaphysics of Morals* (1785), in *Practical Philosophy*, ed. and trans. M. J. Gregor (Cambridge: Cambridge University Press, 1996).
10. Thomas Hobbes, *Leviathan* (1651), ed. R. Tuck (Cambridge: Cambridge University Press, 1996); John Rawls, *A Theory of Justice (*Cambridge, MA: Harvard University Press, 1971); Thomas M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998).
11. For example, Carolyn Raffensperger and Joel Tickner, eds., *Protecting Public Health and the Environment* (Washington, DC: Island Press, 1999); Christian Munthe, *The Price of Precaution and the Ethics of Risk* (Dordrecht: Springer, 2011).
12. For example, Ted Lockhart, *Moral Uncertainty and Its Consequences* (Oxford: Oxford University Press, 2000); Michael J. Zimmerman, *Ignorance and Moral Obligation* (Oxford: Oxford University Press, 2014); William MacAskill, Krister Bykvist, and Today Ord, "Moral Uncertainty," unpublished manuscript, 2018.
13. We include the potentiality condition so as to avoid taking a controversial stand on fetuses and people in comas. Kate Darling, Daniel Estrada, and Greg Antill have argued (each on different grounds) that AI need not even be potentially conscious to deserve moral consideration. Kate Darling, "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior toward Robotic Objects," in *Robot Law,* ed. Ryan Calo, A. Michael Froomkin, and Ian Kerr (Glos, UK: Edward Elgar, 2016); Daniel Estrada, "Robot Rights: Cheap, Yo!," *Made of Robots,* episode 1. May 24, 2017, https://www.youtube.com/watch?v=TUMIxBnVsGc. We have some sympathy with these arguments but will not address them here. See Eric Schwitzgebel, "The Social-Role Defense of Robot Rights." Blog post at *The Splintered Mind* (blog), (June 1, 2017). URL: http://schwitzsplinters.blogspot.com/2017/06/the-social-role-defense-of-robot-rights.html.
14. For a recent deontological view that is explicit about the need of consciousness for moral considerability, see Christine M. Korsgaard, "On Having a Good," *Philosophy* 89 (2014): 405–29.
15. Philip Goff, *Consciousness and Fundamental Reality (*New York: Oxford University Press, 2017); Galen Strawson, *Consciousness and Its Place in Nature (*Exeter: Imprint Academic, 2006).
16. For skepticism about attribution of phenomenal consciousness to infants and nonhuman animals, see Daniel C. Dennett, *Kinds of Minds (*New York: Basic Books, 1996); Peter Carruthers, *Phenomenal Consciousness (*Cambridge: Cambridge University Press, 2000).
17. Thomas Nagel, "What Is It Like to Be a Bat?," *Philosophical Review* 83 (1979): 435–50; Colin McGinn, "Can We Solve the Mind-Body Problem?," *Mind* 98 (1989): 349–66; Ned Block, "The Harder Problem of Consciousness" (2002), in *Consciousness, Function, and Representation* (Cambridge, MA: MIT Press, 2007); Eric Schwitzgebel,

"The Crazyist Metaphysics of Mind," *Australasian Journal of Philosophy* 92 (2014): 665–82.

18. Joanna Bryson advocates one half of the Policy of Excluded Middle: create only robots whose lack of full moral status is clear, so that we are not tempted to give them undeserved rights. Joanna J. Bryson, "Patiency Is Not a Virtue: Intelligent Artifacts and the Design of Ethical Systems," *Ethics and Information Technology* 20 (2018), 15–26 and "Robots Should Be Slaves," in *Close Engagements with Artificial Companions,* ed. Yorick Wilks (Amsterdam: John Benjamins, 2010). We believe this is sensible advice. However, Bryson sometimes seems to encourage robot "slavery," which we think is an unhelpful way of phrasing her point. As Darling, "Extending Legal Protection to Social Robots," and Estrada, "Robot Rights," argue, in virtue of their social roles and our natural psychological responses to them, it might be ethically inappropriate to treat some socially important robots in ways we associate with slavery.
19. John Basl, "The Ethics of Creating Artificial Consciousness," *APA Newsletter on Philosophy and Computers* 13, no. 1 (2013): 23–29; John Basl, "Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines," *Philosophy & Technology* 27 (2014): 79–96.
20. Inspired by Douglas Adams, "The Restaurant at the End of the Universe," in *The Ultimate Hitchhiker's Guide to the Galaxy* (1980; New York: Random House, 2002).
21. Compare Mary Poppins 3000 in Mark Walker, "A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!," in *Human Implications of Human-Robot Interaction*, ed. T. Metzler (AAAI Press, 2006), http://dept-wp.nmsu.edu/philosophy/files/2014/07/ws0610walkera.pdf.
22. Steve Petersen, "The Ethics of Robot Servitude," *Journal of Experimental and Theoretical Artificial Intelligence* 19 (2007): 43–54; Steve Petersen, "Designing People to Serve," in *Robot Ethics,* ed. P. Lin, K. Abney, and G. A. Bekey (Cambridge, MA: MIT Press, 2012).
23. See, for example, Jonathan Glover, *Choosing Children* (Oxford: Oxford University Press, 2006); Allen E. Buchanan, *Beyond Humanity?* (Oxford: Oxford University Press, 2011); Robert Sparrow, "A Not-So-New Eugenics," *Hastings Center Report* 41, no. 1 (2011): 32–42; Sara Goering, "Eugenics," *Stanford Encyclopedia of Philosophy,* July 2, 2014, https://plato.stanford.edu/archives/fall2014/entries/eugenics/.
24. Maybe there are aesthetic goals so valuable that one might reasonably enough choose to sacrifice one's life for them. If necessary, we can stipulate that the Cow at the End of the Universe is not a case like that. The Cow is no great aesthete, and it knows that it will become an about-average set of steaks in a mundane, forgettable aesthetic experience for the jaded restaurant patrons.
25. On life expectancy by occupation in eighteenth-century Berlin, see Helga Schulz, "Social Differences in Mortality in the Eighteenth Century: An Analysis of Berlin Church Registers," *International Review of Social History* 36 (1991): 232–48; on indentured white servants in the colonial United States, see Don Jordan and Michael Walsh, *White Cargo (*New York: New York University Press, 2007); and on the relative life expectancies of masters and servants in nineteenth-century Britain, see Lucy

Lethbridge, *Servants: A Downstairs History of Britain from the Nineteenth Century to Modern Times* (New York: Norton, 2013).

26. Compare Justin White, "Why Did the Butler Do It? Autonomy, Authenticity, and Human Agency," unpublished manuscript, 2018, on the difficulties faced by the butler Stevens in Kazuo Ishiguro's *Remains of the Day* (New York: Vintage, 1988), in maintaining dignity and autonomy given his extreme deference to his employer. On the challenges of autonomy in deferential roles, see also Andrea C. Westlund, "Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy?," *Philosophical Review* 112 (2003): 483–523; Marina Oshana, *Personal Autonomy in Society* (Hampshire, UK: Ashgate, 2006); James Rocha, "Autonomy within Subservient Careers," *Ethical Theory and Moral Practice* 14 (2011): 313–28. (P. G. Wodehouse's Jeeves, for the record, is quite capable of forming independent autonomous plans into which he steers Wooster.)
27. Compare McKenna's Suzie Instant and Mele's "minuteling": Michael McKenna, "A Modest Historical Theory of Moral Responsibility," *Journal of Ethics* 20 (2016): 83–105; Michael McKenna, "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32 (2004): 169–82; Alfred R. Mele, "Moral Responsibility, Manipulation, and Minutelings," *Journal of Ethics* 17 (2013): 153–66. One difference is that McKenna's Suzie Instant and Mele's minuteling have false memories and Sun Probe has no false memories. McKenna's "positive historical" thesis is that freedom and moral responsibility require one's actions arise from values that one has had an opportunity to critically assess. Compare also John Martin Fischer and Mark Ravizza, *Responsibility and Control* (Cambridge: Cambridge University Press, 1998). Our thesis doesn't require that Sun Probe has *no* freedom, responsibility, or autonomy, only that its freedom, responsibility, or autonomy is impaired and that it deserves a developmentally extended opportunity to explore and possibly alter its values.

    We favor a "compatibilist" view of freedom on which freedom in the relevant sense is compatible with determinism. However, we hope that the argument here can be reconciled with libertarian views (if Second Probe can be endowed with whatever metaphysical free will biological human beings have) and with hard determinist views (if we hold Second Probe to the same types of standards we hold ourselves, despite lack of freedom).
28. Should this opportunity include the opportunity not only to settle on values not just somewhat at variance with our own but also, possibly, to settle on values that are radically morally abhorrent? This is a tricky question in human cases also. To what extent should parents or societies forbid people from exploring, for example, Nazi values, as opposed to only strongly discouraging such values and trusting that reasonable people in free discussion will come to reject them?
29. According to free will theodicy, God faced essentially the same choice in creating humans. In Schwitzgebel and Garza, "A Defense of the Rights of Artificial Intelligences," we argue that in some AI creation scenarios the designers would literally be gods relative to the AIs, with the moral responsibilities pertaining thereto.
30. For valuable discussion and comments, thanks to Greg Antill, Daniel Estrada, John Fischer, Steve Petersen, and Eli Rubinstein; audiences at New York University and UCLA; and the many commenters on relevant posts at *The Splintered Mind* and Eric Schwitzgebel's Facebook page.

References

Adams, Douglas. "The Restaurant at the End of the Universe." In *The Ultimate Hitchhiker's Guide to the Galaxy*. 1980; New York: Random House, 2002.

Antill, Gregory E. "Robots and Reactive Attitudes: In Defense of the Moral and Interpersonal Status of Non-conscious Creatures." Unpublished manuscript , 2018.

Basl, John. "The Ethics of Creating Artificial Consciousness." *APA Newsletter on Philosophy and Computers* 13, no. 1 (2013): 23–29.

Basl, John. "Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines." *Philosophy & Technology* 27 (2014): 79–96.

Bentham, Jeremy. *The Principles of Morals and Legislation*. 1789; Amherst, NY: Prometheus, 1988.

Block, Ned. "The Harder Problem of Consciousness" (2002). In *Consciousness, Function, and Representation*. Cambridge, MA: MIT Press, 2007.

Briggs, Rachael, and Daniel Nolan. "Utility Monsters for the Fission Age." *Pacific Philosophical Quarterly* 96 (2015): 392–407.

Brink, David O.. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press, 1989.

Bryson, Joanna J. "Patiency Is Not a Virtue: Intelligent Artifacts and the Design of Ethical Systems," *Ethics and Information Technology* 20 (2018), 15–26.

Bryson, Joanna J. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions*, edited by Yorick Wilks. Amsterdam: John Benjamins, 2010.

Buchanan, Allen E. *Beyond Humanity?* Oxford: Oxford University Press, 2011.

Carruthers, Peter. *Phenomenal Consciousness.* Cambridge: Cambridge University Press, 2000.

Coeckelbergh, Mark. *Growing Moral Relations*. Basingstoke, UK: Palgrave Macmillan, 2012.

Darling, Kate. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior toward Robotic Objects." In *Robot Law,* edited by Ryan Calo, A. Michael Froomkin, and Ian Kerr. Glos, UK: Edward Elgar, 2016.

Dennett, Daniel C. *Kinds of Minds*. New York: Basic Books, 1996.

Dick, Philip K. *Do Androids Dream of Electric Sheep?* New York: Doubleday, 1968.

Estrada, Daniel "Robot Rights: Cheap, Yo!" *Made of Robots,* episode 1. May 24, 2017. https://www.youtube.com/watch?v=TUMIxBnVsGc.

Fancher, Hampton, Michael Green, and Denis Villeneuve. *Blade Runner 2014*. Burbank, CA: Warner Bros., 2017.

Fancher, Hampton, David Peoples, and Ridley Scott. *Blade Runner.* Burbank, CA: Warner Bros., 1982.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control.* Cambridge: Cambridge University Press, 1998.

Flanagan, Owen, Hagop Sarkissian, and David Wong. "Naturalizing Ethics." In *Moral Psychology,* vol. 1, edited by Walter Sinnott-Armstrong. Cambridge, MA: MIT Press, 2007.

Glover, Jonathan. *Choosing Children*. Oxford: Oxford University Press, 2006.

Goering, Sara. "Eugenics." *Stanford Encyclopedia of Philosophy*, July 2, 2014. https://plato.stanford.edu/archives/fall2014/entries/eugenics/.

Goff, Philip. *Consciousness and Fundamental Reality.* New York: Oxford University Press, 2017.

Gunkel, David J. *The Machine Question.* Cambridge, MA: MIT Press, 2012.
Hobbes, Thomas. *Leviathan* (1651). Edited by R. Tuck. Cambridge: Cambridge University Press, 1996.
Ishiguro, Kazuo. *The Remains of the Day*. New York: Vintage, 1988.
Jordan, Don, and Michael Walsh. *White Cargo.* New York: New York University Press, 2007.
Kant, Immanuel. *Critique of Practical Reason* (1788). In *Practical Philosophy*, edited and translated by M. J. Gregor. Cambridge: Cambridge University Press, 1996.
Kant, Immanuel. *Groundwork of the Metaphysics of Morals* (1785*).* In *Practical Philosophy*, edited and translated by M. J. Gregor. Cambridge: Cambridge University Press, 1996.
Korsgaard, Christine M. "On Having a Good." *Philosophy* 89 (2014): 405–29.
Lethbridge, Lucy. *Servants: A Downstairs History of Britain from the Nineteenth Century to Modern Times*. New York: Norton, 2013.
Lockhart, Ted. *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press, 2000.
MacAskill, William, Krister Bykvist, and Toby Ord. "Moral Uncertainty." Unpublished manuscript, 2018.
McGinn, Colin. "Can We Solve the Mind-Body Problem?" *Mind* 98 (1989): 349–66.
McKenna, Michael. "A Modest Historical Theory of Moral Responsibility." *Journal of Ethics* 20 (2016): 83–105.
McKenna, Michael. "Responsibility and Globally Manipulated Agents." *Philosophical Topics* 32 (2004): 169–82.
Mele, Alfred R. "Moral Responsibility, Manipulation, and Minutelings." *Journal of Ethics* 17 (2013): 153–66.
Mill, John Stuart. *Utilitarianism* (1861). Edited by G. Sher. Indianapolis, IN: Hackett, 2001.
Munthe, Christian. *The Price of Precaution and the Ethics of Risk*. Dordrecht: Springer, 2011.
Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (1979): 435–50.
Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.
Oshana, Marina. *Personal Autonomy in Society*. Hampshire, UK: Ashgate, 2006.
Petersen, Steve. "The Ethics of Robot Servitude." *Journal of Experimental and Theoretical Artificial Intelligence* 19 (2007): 43–54.
Petersen, Steve. "Designing People to Serve." In *Robot Ethics,* edited by P. Lin, K. Abney, and G. A. Bekey. Cambridge, MA: MIT Press, 2012.
Raffensperger, Carolyn, and Joel Tickner, eds. *Protecting Public Health and the Environment.* Washington, DC: Island Press, 1999.
Railton, Peter. "Moral Realism." *Philosophical Review* 95 (1986): 163–207.
Rawls, John. *A Theory of Justice.* Cambridge, MA: Harvard University Press, 1971.
Rocha, James. "Autonomy within Subservient Careers." *Ethical Theory and Moral Practice* 14 (2011): 313–28.
Scanlon, Thomas M. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.
Schulz, Helga. "Social Differences in Mortality in the Eighteenth Century: An Analysis of Berlin Church Registers." *International Review of Social History* 36 (1991): 232–48.
Schwitzgebel, Eric. "The Crazyist Metaphysics of Mind." *Australasian Journal of Philosophy* 92 (2014): 665–82.
Schwitzgebel, Eric "The Social-Role Defense of Robot Rights." *The Splintered Mind* (blog), June 1, 2017. http://schwitzsplinters.blogspot.com/2017/06/the-social-role-defense-of-robot-rights.html.

Schwitzgebel, Eric, and Mara Garza. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39 (2015): 98–119.

Searle, John R. (1980). "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–57.

Searle, John R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press, 1992.

Sparrow, Robert. "A Not-So-New Eugenics." *Hastings Center Report* 41, no. 1 (2011): 32–42.

Tononi, Giulio. "The Integrated Information Theory of Consciousness: An Updated Account." *Archives Italiennes de Biologie* 150 (2012): 290–326.

Walker, Mark. "A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!" In *Human Implications of Human-Robot Interaction*, edited by T. Metzler. AAAI Press, 2006. http://dept-wp.nmsu.edu/philosophy/files/2014/07/ws0610walkera.pdf.

Westlund, Andrea C. (2003). "Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy?" *Philosophical Review* 112 (2003): 483–523.

White, Justin "Why Did the Butler Do It? Autonomy, Authenticity, and Human Agency." Unpublished manuscript, 2018.

Zimmerman, Michael J. (2014). *Ignorance and Moral Obligation*. Oxford: Oxford University Press, 2014.

17
The Moral Status and Rights of Artificial Intelligence

S. Matthew Liao

17.1. Introduction

Artificial intelligence (AI) is becoming more and more capable. In 2011 IBM's Watson defeated two of the best human players on *Jeopardy!*, Ken Jennings and Brad Rutter.[1] Prior to their defeat, Jennings had been unbeaten in seventy-four appearances and Rutter had earned a total of $3.25 million on the show. In 2016 Google DeepMind's AlphaGo played five games of Go against the eighteen-time world champion, Lee Sedol, and AlphaGo won four out of five games.[2] This marked the first time a machine had defeated the world's best player at this ancient game. Indeed, experts thought that a machine would not be able to do so for some time, owing to the complexity of the game and the fact that it seems impossible for current computers to use brute force in order to search all the possible moves and find the best move. In 2017 AlphaGo Zero, a more sophisticated version of AlphaGo that learned to play Go by playing against itself instead of using data from human games, beat AlphaGo 100 games to 0.[3] The algorithms used to train AlphaGo Zero were then adapted to play chess, and this version, called AlphaZero, was able to beat one of the top chess programs in the world, Stockfish, 28 times, with 0 losses and 72 draws.[4] And efforts are underway to develop increasingly advanced AI in other domains. For instance, several companies, including Microsoft, Google, and Affectiva, are using facial recognition, voice recognition, and deep learning to build AIs that can sense and respond to facial expressions of emotions.[5] Recently, Boston Dynamics has demonstrated that its robot dog can open doors on its own.[6]

As AIs acquire greater capacities, the question of whether AIs will acquire greater moral status becomes salient. Moral status is the standing an entity has that gives moral agents a pro tanto reason to act toward it in a certain way.[7] For example, human beings are rightsholders. As such, a moral agent is, for example, prohibited from killing a human being just for personal benefit. Likewise cats have a certain moral status that gives moral agents at least a pro tanto reason to act in a certain way toward them. In particular, moral agents have a duty not to

S. Matthew Liao, *The Moral Status and Rights of Artificial Intelligence* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0018

inflict pain on cats just for fun because cats are sentient. In 2017 Saudi Arabia granted citizenship to a robot from Hanson Robotics named Sophia.[8] Also, the European Parliament has argued in favor of giving "electronic personhood" and legal rights to certain AIs.[9] Given these developments, it seems appropriate to consider the kinds of moral status that AIs could have as they acquire greater capacities.[10] In particular, could AIs achieve human-level moral status and be rightsholders? If AIs could be rightsholders, what rights would they have? Could AIs have greater than human-level moral status? The goal of this chapter is to shed light on some of these questions. To do so, I begin by sketching a theory of moral status and considering what kind of moral status an AI can have.

17.2. Moral Status and AI

To start, it is useful to have an idea of who has or could have moral status.[11] Here is a nonexhaustive list of different entities that have or could have moral status:

1. Inanimate objects (rocks, artworks, buildings, the environment).
2. Nonhuman terrestrial living things (plants and animals).
3. Normal-functioning human beings with full physical, cognitive, emotional, and social capacities (normal adult human beings).
4. Injured human beings (the comatose, the severely mentally disabled).
5. Human beings at the beginning of life (fetuses, infants, young children).
6. Possible or future human beings (future generations).
7. Nonliving human beings (dead human beings).
8. Nonhuman extraterrestrial species of living beings—should they exist (alien beings from outer space).
9. Artificial life forms (androids; robots; computers; algorithms).

How do we determine what kind of moral status each of these entities has? There are some helpful constraints to guide this inquiry. First, we need to know what kind of empirical attributes a particular entity has. The reason for this constraint is that there does not seem to be a purely a priori way of knowing what kind of moral status an entity has. For example, suppose that we encounter an alien being from outer space for the first time, and we want to know what kind of moral status the alien has. It seems that we would not be able to know this alien's moral status solely through some a priori way. To determine the moral status of this alien being, it seems that we would at least have to investigate empirically what attributes this alien being has and decide whether the empirical attributes that it has warrant according this alien a certain kind of moral status.

Second, any proposed empirical attribute or criterion for moral status should meet the Species Neutrality Requirement and be nonspeciesist, where speciesism is defined as morally favoring a particular species over others without sufficient justification. The Species Neutrality Requirement says that a proper criterion for moral status should not exclude any species in advance. For instance, suppose that one proposes that "being human" is necessary for being a rightsholder. This criterion would appear to exclude all nonhuman species from having rights in advance. As such, "being human" would not be an appropriate criterion for moral status.

Third, and not uncontroversially, I submit that the moral status of an entity should be based on the intrinsic properties of that entity. These are properties that are internal or inherent to an entity. The extrinsic properties of an entity are those that an entity has in virtue of its relationship with other entities. For example, being a moral agent is an intrinsic property that a normal-functioning human being typically has. Being a spouse, on the other hand, is an extrinsic property that someone has in virtue of being married to another person. A reason for thinking that moral status should be based on intrinsic rather than extrinsic properties is to imagine two entities having exactly the same intrinsic properties. It seems that other things being equal, the two entities should have the same moral status regardless of the relationships these entities may have with others. This is not to say that extrinsic properties are entirely irrelevant, but only that they should not affect an entity's moral status.

To illustrate these points, suppose that your spouse and a stranger are both drowning and only one of them can be saved. It seems morally permissible for you to choose to save your spouse. Does this imply that your spouse has a higher moral status than the stranger in virtue of her relationship with you? If it did, it would seem to imply that all agents, and not just you, have a stronger pro tanto reason to save your spouse before the stranger.[12] However, it seems more plausible to regard your spouse and the stranger as having the same moral status (in virtue of both possessing similar intrinsic properties such as moral agency), and to regard the extrinsic property of being a spouse as a tie-breaker. In other words, it seems that extrinsic, relational properties such as "is the spouse of" give only some agents a stronger pro tanto reason to promote an individual's interests. If so, it seems that only intrinsic properties are of the sort that can ground moral status.[13]

Suppose that this is correct. What are some empirical, nonspeciesist, intrinsic properties that an entity could have that would give it a certain moral status? Here are some candidates:

- Being alive.
- Being conscious.
- Being able to feel pain.

- Being able to desire.
- Being capable of rational agency, for example, being able to know something about causality such as if one does x, then y would happen, and being able to bring about something intentionally.
- Being capable of moral agency, such as being able to understand and act in light of moral reasons.

For instance, in virtue of being alive, plants arguably have a certain moral status. Other things being equal, moral agents should not destroy a plant for no reason whatsoever. Likewise, in virtue of being able to feel pain, turtles have a certain moral status. Other things being equal, moral agents should not cause a turtle to suffer for no reason.

Next, it seems that some entities have a greater moral status than others. X has a greater moral status than Y if and only if X deserves greater respect and/or protection than Y, other things being equal. For instance, if one could save either X or Y but not both, that X deserves greater respect and/or protection than Y implies that one should, all else being equal, save X. Compare a rock and a plant. It seems that the plant would have a greater moral status, other things being equal, owing to the fact that the plant is alive but the rock is not. In other words, a plant deserves greater respect and/or protection than a rock, other things being equal. Of course, other things are not always equal. Suppose that a rock has a certain extrinsic value (it is the diamond from your engagement ring), while a plant has a certain extrinsic disvalue (it is a weed in your backyard that you are trying to get rid of). In this case, one might not think that the plant would deserve greater respect and/or protection than the rock. Still, recall that moral status depends solely on an individual's intrinsic properties. From this perspective, other things being equal, there are good reasons to think that a living entity should have a greater moral status than a nonliving one.

Among living entities, some also arguably have a greater moral status than others. Compare a plant and a turtle. All else being equal, it seems that a turtle has a greater moral status than a plant, owing to the fact that a turtle can feel pain while a plant cannot. Finally, among entities capable of feeling pain, arguably, some have even greater moral status than others. Compare a normally functioning adult human being and a turtle. It seems that a human being has a far greater moral status than a turtle. An explanation is that the human being has moral agency while the turtle does not.

In light of this discussion, the kind of moral status that an AI can have will depend on the kinds of empirical, nonspeciesist, intrinsic properties that an AI has, setting aside epistemic issues about how we can know whether an AI has these properties and bracketing debates (in, e.g., functionalism) about whether AIs are the kind of things that can have these properties. For instance, is the AI

alive, conscious, and/or sentient? Is the AI able to feel pain? Is the AI capable of having desires? Does the AI have rational and/or moral agency? Supposing that the AI acquires some of these intrinsic properties, it seems that the AI should have the same kind of moral status as other entities that have the same intrinsic properties. For instance, if an AI is alive, then the AI should have moral status at least on par with living things such as plants. If an AI can feel pain, then the AI should have moral status at least on par with beings capable of feeling pain, such as turtles. To decide whether an AI could have human-level moral status, we need to discuss the kind of empirical, nonspeciesist, intrinsic property that grounds human moral status, that is, rightsholding. I turn to this issue now.

17.3. AI and Rightsholding

There is an intuitive thought that all human beings are rightsholders.[14] Indeed, the Universal Declaration of Human Rights (UDHR, 1948) explicitly states that "all human beings are born free and equal in dignity and rights." However, it turns out that showing that all human beings are rightsholders is complicated. In fact, philosophers who have examined this issue have often find themselves either accepting that not all human beings can be rightsholders or adopting what Peter Singer has called a "speciesist" position, which morally favors human beings over other species without sufficient justification. This is because philosophers have found it difficult to identify a relevant empirical, nonspeciesist, intrinsic property that applies to all human beings. For instance, consider actual sentience. Some human beings such as anencephalic children and comatose persons lack actual sentience. Or, consider actual moral agency. Many human beings, including newborn infants, lack actual moral agency. If human moral status were grounded in these properties, these human beings would not be rightsholders.

Elsewhere I have argued that we can overcome this impasse by noticing that there is an empirical, nonspeciesist, intrinsic property that seems to apply to all human beings.[15] In particular, I have proposed that a sufficient condition for being a rightsholder is having the genetic (or, more generally, the physical) basis for moral agency, and that all human beings have the genetic basis for moral agency. Let me briefly explicate this account.

By the physical, genetic basis for moral agency, I mean the set of physical codes that generate moral agency. In human beings, this set of codes is located in their genome. We know this because the developmental basis for adaptive phenotypes like moral agency requires a great deal of complexity, and the genome contains a significant proportion of this complexity. At present we do not know exactly which set of genes is necessary and sufficient for the genetic basis for moral agency. But we can talk about a genetic basis for moral agency as long as there are

genes that definitely play no role in forming the genetic basis for moral agency. For example, the genes for my toenails or a gene whose expression only produces pigment in the eyes probably plays no role in the formation of the genetic basis for moral agency. Also, to have the genetic basis for moral agency, the genes that make up moral agency must be activated and be coordinating with each other in an appropriate way. To illustrate the point about coordination, consider the following: Suppose there is a book containing many random words which, if put together in the right way, would result in a Shakespeare book. That book would not be a Shakespeare book just because it contained the correct words; those words must be organized in the right way.

There are reasons to believe that all human beings have the genetic basis for moral agency. For instance, we know that all normal-functioning and normal-developing human beings have this genetic basis because they exercise moral agency or will exercise it. We also know that most comatose human beings have this genetic basis because they exercised moral agency in the past. Moreover, we know that human beings with mild mental retardation, such as children with Down syndrome, typically exhibit some moral agency, which suggests that they also have this genetic basis. Finally, to see how human beings with severe disabilities that are the result of genetic defects rather than environmental factors would also have this genetic basis, it is useful to distinguish between genetic defects of the genes that make up an attribute and genetic defects that undermine the development of an attribute. Consider a human being born without a hand. This may be because this human being lacks the genes to form the hand, or it may be that certain conditions needed for the genes to form the hand, such as prenatal nutrition, were blocked or lacking. In the former, this human being would not have the genetic basis for having a hand, since the human being lacks the genes that make up the hand. In the latter, the human being would still have the genetic basis for having a hand, because the genes that make up the hand are present and active, but they were blocked from developing owing to certain conditions.

On this distinction, the genetic defects in human beings with severe disabilities do not seem to be defects in the genetic basis for moral agency but at best defects that undermine the development for moral agency. Consider phenylketonuria (PKU), Tay-Sachs, Sandhoff disease, and a whole cluster of about seven thousand other kinds of genetic disorders, which are caused by the mutation of a gene.[16] The gene is typically necessary for producing a certain protein or enzyme, which is then needed to change certain chemicals to other ones or to carry substances from one place to another. Mental retardation and other defects are typically caused by abnormal buildups of certain amino acids that become toxic to the brain and other tissues, because the cell is unable to process these amino acids owing to the mutation. But with treatment of a low-enzyme diet as soon as possible in the neonatal stage, normal growth and cognitive development can be

expected in many cases. This shows that the brain tissue initially developed properly and would have continued to do so if excessive amino acids had not accumulated. Therefore, following the distinction between genetic defects that make up an attribute and genetic defects that undermine the development of the attribute, single-gene defects seem to be cases of the latter rather than the former. Given this, human beings who have these kinds of genetic defects most likely have the genetic basis for moral agency.[17]

Let me mention some virtues of this account of rightsholding and take up some objections. One advantage of this account is that the physical, genetic basis for moral agency is an identifiable, empirical, intrinsic property. Another virtue of this account is that it meets the Species Neutrality Requirement and is therefore not speciesist. Indeed, on this account, if we were to learn that chimpanzees or some other animals have the genetic basis for moral agency, then they too would be rightsholders.

Now some people might think that moral agency matters only if one can actually exercise it. On this view, the possession of the physical, genetic basis for moral agency does not matter. What matters is that one actually has the capacity to act in light of moral reasons. Indeed some people will claim that the value of the genetic basis for moral agency is entirely derived from the value of actual moral agency. However, actual moral agency cannot be the sole ground for rightsholding. The reason is that if rightsholding serves any function at all, one function that it would serve would be the following:

> If and when the rightsholder's interest is in conflict with the same kind of interest, that is, with the comparable interest, of a non-rightsholder, the rightsholder's interest should prevail.

A corollary of this is that if one were to give the interest of a non-rightsholder priority over the comparable interest of a rightsholder, then one would be acting wrongly. For example, normal-functioning adult human beings are typically regarded as rightsholders, whereas normal adult turtles are not. (Those who believe that all animals have rights can substitute the turtle with whatever they would regard as a non-rightsholder. The general point would remain valid.) Suppose that this is correct, and suppose rightsholding has the function I suggested. Consider a case where one can either save a turtle's limb or a human being's limb. Since the turtle's interest in keeping the limb and the human being's interest in keeping the limb appear to be comparable, if rightsholding has the function I suggested, it seems that one should save the human being's limb. If one does not save the human's limb, then it seems that one would have acted wrongly. Note that this does not mean that one has no duties regarding non-rightsholders. For example, if a turtle is about to lose a limb and saving it requires

little effort, it seems that one would have a duty to do so. It also does not mean that *any* interest of a rightsholder is necessarily more important than any interest of a non-rightsholder. If the turtle is about to lose a limb and I, a rightsholder, would merely have to let my cup of tea become cold to save the turtle's limb, it seems that I may have a duty to do so.

If all of this is true, we should be able to see why actual moral agency cannot be the sole ground for rightsholding. If it were, it would imply that adult human beings are rightsholders while human infants are not, as the latter do not have actual moral agency. If rightsholding has the function I suggested, it would mean that an adult human being's interest should be prioritized over the comparable interest of an infant. And if one were to give an infant's interest priority over the comparable interest of an adult human being, then one would be acting wrongly. Yet it is often permissible to prioritize the interest of infants over the comparable interest of adult human beings. For example, suppose a human baby with many years of life left and a human adult who has only a short period left to live are drowning, and one can save only one of them. It seems permissible to save the infant rather than the adult. Or, if a ship is sinking, it seems permissible to give infants the priority to be in the lifeboat instead of the adults.[18] These examples suggest that either we are wrong to think that it is sometimes permissible to give infants preference over adults, or that actual moral agency is not the sole basis for rightsholding. Our judgment is that it is at least permissible and not morally wrong to prioritize infants over adults. If so, this suggests that actual moral agency cannot be the sole ground for rightsholding.

Suppose that having the genetic (or, more generally, the physical) basis for (the development of) moral agency is in fact a sufficient condition for being a rightsholder. This gives us one way to determine when an AI would have human-level moral status and be a rightsholder. In particular, an AI would be a rightsholder if it has the physical basis for (the development of) moral agency. Similar to human beings, the physical basis for moral agency for an AI could be the set of physical codes that generate moral agency in the AI. As with the genetic basis for moral agency, the physical basis for moral agency should be activated and be coordinating with each other in an appropriate way. An AI would lack the physical basis for moral agency if it possessed the algorithms necessary for moral agency, but those algorithms either were not activated or were scrambled in such a way that they do not coordinate with each other in an appropriate way.

Could AIs acquire the physical basis for moral agency in the future even though they almost certainly do not have it now? I shall consider three possible ways by which they may come to acquire the physical basis for moral agency. One possibility is through "mind uploading," that is, some kind of brain emulation.[19] For instance, one might be able to perform a high-resolution scan of a brain and capture all the neurons and their synaptic interconnections.[20] One might then be

able to construct a computational software analog of these neurons and synaptic interconnections such that if the software analog were to run on appropriate hardware, it would exhibit the essential functional characteristics of the original brain. One could then host the emulated AI in a simulation or place the emulated AI in a robotic body so that it could interact with the external world. Suppose that the emulated AI exhibits actual moral agency. One might then infer that the emulated AI has the physical basis for moral agency.

Although brain emulation is not currently feasible, it is worth noting that many research groups around the world are actively pursuing projects that could enable this kind of procedure in the future. For instance, the European Union has invested over a billion euros into the Human Brain Project, which aims to emulate the brain of a mouse and parts of the human brain by 2023.[21] In the United States, the BRAIN Initiative and the Human Connectome Project are trying to map the neural pathways that underlie human brain function in order to acquire and share data about the structural and functional connectivity of the human brain.[22] Recently a start-up company called Nectome has offered people with terminal illnesses the possibility of preserving their brains using a high-tech embalming process so that their brains could be emulated at a future date.[23]

A second way by which an AI might be able to acquire the physical basis for moral agency is a process of gradual substitution.[24] Gradual substitution involves replacing the carbon-based cells in an individual's brain gradually, bit by bit, with functionally equivalent inorganic substitutes until all the carbon-based cells are replaced with the inorganic substitutes. Suppose that life-sustaining processes such as absorption, assimilation, and metabolism are maintained throughout this procedure, and suppose that the cerebrum continues to be activated in the normal way such that consciousness is not interrupted throughout this procedure. The thought is that if the individual had the physical basis for moral agency before gradual substitution, the resulting individual should also have the physical basis for moral agency after gradual substitution.

A third way by which an AI could have the physical basis for moral agency is to build such an AI through computer programming and algorithms. How could this be achieved? Some people have entertained the possibility of compiling all moral rules and specifying them in algorithms for an AI to follow.[25] Others have suggested hard-coding general, top-down moral theories such as consequentialism, deontology, and virtue ethics into AIs so that AIs can use them to determine what the morally right action is on a particular occasion.[26] However, it is not clear that moral agency can be built into AIs in such ways. For one thing, if moral rules vary according to circumstances and context, then there could be an infinite number of moral rules. If so, it would be a tall task to try to build them into AIs. In addition, general, top-down moral theories such as consequentialism, deontology, and virtue ethics seem too specific since not everyone

accepts these theories. Indeed there continue to be lively debates regarding the veracity of these theories. Given this, it does not seem like the right approach to hard-code something as specific and debatable as consequentialism or deontology or virtue ethics into an AI. Most important, though, there is an issue of whether AIs built in these ways could grasp and understand why a certain action is morally right or wrong. If these AIs could only follow moral rules but not understand the rationale or grounds for them, they would not seem to have moral agency.[27] A fortiori, they would not have the physical basis for moral agency.

Here may be a more promising, bottom-up, way by which one may build the physical basis for moral agency into an AI. In linguistics, some people have proposed that the mind may be equipped with a universal set of principles or "grammar" that enables any normally developing human being in different cultures unconsciously to generate and comprehend a limitless range of well-formed sentences in his or her native language.[28] Drawing on this linguistic analogy, some people have proposed that the mind may also be equipped with a universal "moral" grammar that enables each of us unconsciously and automatically to evaluate a limitless variety of actions and generate moral evaluations such as right and wrong.[29] According to one version of this universal moral grammar hypothesis, the moral faculty works as follows. There are domain-general cognitive mechanisms that generate representations of actions in terms of variables such as AGENT, INTENTION, BELIEF, ACTION, RECEIVER, CONSEQUENCE, MORAL EVALUATION.[30] Some cognitive mechanisms that comprise the moral faculty then combine these representations to generate moral judgments such as "impermissible," "permissible," and "obligatory."

We might regard these domain-general cognitive mechanisms as providing the basic components of our moral agency. As such, we might try to build these cognitive mechanisms into AIs, thereby creating some kind of Artificial Moral Grammar. To do so, we would need to develop computer algorithms that could represent action in terms of variables such as AGENT, INTENTION, BELIEF, ACTION, RECEIVER, CONSEQUENCE, MORAL EVALUATION. We would also need to develop algorithms that could combine these representations to generate moral judgments, such as "impermissible," "permissible," and "obligatory."

How would we know whether or not we succeeded in creating genuine Artificial Moral Intelligence (AMI)? As suggested earlier, it seems that we will have succeeded when we have created an AI that can grasp and understand the reasons why a certain action is morally right or wrong. For instance, an AMI would not just know that lying is typically wrong, but would also understand that lying is typically wrong because it fails to treat people respectfully. Once a genuine AMI has been created, we can infer that it would have the physical basis for moral agency.

Among the three possible ways of creating an AI with the physical basis for moral agency, should we prefer one particular way? The answer to this question is likely to depend on our goals and objectives when creating such an AMI. To give an example, in the literature on creating intelligent AIs, some people have argued that it may be important for humans themselves to become and survive as AIs someday.[31] Suppose that this is the goal. And suppose further that an individual cares about preserving her numerical (and not just qualitative) identity in the process of becoming an AI. The distinction between qualitative and numerical identity is that two things can share certain properties and thus be qualitatively identical without being numerically identical.[32] For instance, two cars can be the same brand and model, be produced in the same year, be from the same factory, have the same color, and so on, and therefore be qualitatively identical. They would, however, not be numerically identical, because numerical identity can hold only between a thing and itself. As an example of numerical identity, I am numerically identical to the individual sitting in front of this computer right now.

A case can be made here that an individual should prefer gradual substitution to the other two methods. Why is this? It should be clear that if an AI were built by computer programming to resemble an individual, that AI would not be that individual, numerically speaking, even if that AI had all of the individual's characteristics, memories, and personality, qualitatively speaking. Psychological theories of personal identity notwithstanding, it seems that an individual whose brain has been emulated and uploaded would also not be numerically identical to the emulation.[33] To see this, suppose that a scientist emulates the functions, structures, and content of my brain, and then runs the emulation on some piece of computer hardware. Call the resulting entity EmuMatthew. Suppose that BioMatthew, who is just me as I am presently constituted, continues to be alive. Would I be numerically identical to EmuMatthew or to BioMatthew? It seems that I would be numerically identical to BioMatthew and not EmuMatthew. In addition, suppose that the scientist then destroys BioMatthew. Since I was not numerically identical to EmuMatthew, it seems that I would not become numerically identical to EmuMatthew now, even though BioMatthew no longer exists. Moreover, suppose that the scientist decides to create and run fifty copies of EmuMatthew on fifty different computers. It should be clear that I would not be numerically identical to all fifty copies of EmuMatthew and that each copy would not be numerically identical to each other. All of this suggests that an individual whose brain has been emulated and uploaded would not be numerically identical to the emulation.

In contrast, there are reasons to think that an individual could survive, numerically speaking, gradual substitution.[34] As discussed earlier, in gradual substitution, life-sustaining processes such absorption, assimilation, and metabolism

are maintained, and the cerebrum continues to be activated in the normal way throughout this procedure. This suggests that functional organismic continuity is maintained through gradual substitution. On an organism view of personal identity, according to which we are essentially organisms, we persist and continue to do so as long as organismic continuity is maintained. If so, at least on an organismic view of personal identity, an individual could survive, numerically speaking, gradual substitution. If so, those who are interested in becoming an AI and who have the goal of surviving, numerically speaking, as an AI should prefer gradual substitution over the other two ways of creating AIs.

17.4. The Rights of AI

Suppose that AIs can achieve human-level moral status and be rightsholders. Call these "rightsholding AIs." What rights would they have? Presumably, they would have some rights that human beings have as well as rights that are unique to them as AIs.[35] Can we offer a systematic, substantive theory that would explain why they have certain rights? To develop such a theory, it might be helpful to consider a similar theory that explains why human beings have the rights that they have. To keep the discussion simple, let us set aside legal rights, which depend in large part on social conventions, and focus on the moral rights that human beings have, in particular, human rights. Elsewhere I have argued that human beings have human rights to the fundamental conditions for pursuing a good life.[36] I shall suggest that this Fundamental Conditions Approach to human rights can help us identify some of the rights that a rightsholding AI could have. To see this, let me first explain what I mean by a fundamental condition for pursuing a good life, why human beings have human rights to these fundamental conditions, and what kind of human rights human beings can have on this account.

As I see it, a good human life is one spent in pursuing certain valuable, basic activities, where basic activities are those that are important to a human being qua human being's life as a whole. For example, sunbathing is an activity but not a basic one, because a human being qua human being's life as a whole is not affected if a human being does not go sunbathing. In addition, activities that are very important to an individual human being's life as a whole may nevertheless not be basic activities, because they may not be important to human beings qua human beings' life as a whole. For instance, being a professional philosopher is very important to my life as a whole. But being a professional philosopher is not a basic activity because it is not an activity that is important to human beings qua human beings' life as a whole. Finally, basic activities are ones such that if a human life did not involve the pursuit of any of them, then that life could not be a good life. An important implication of this point is that a human being can have

a good life by pursuing just some, and not all, of the basic activities. Some of the basic activities are as follows: deep personal relationships with, for instance, one's partner, friends, parents, children; knowledge of, for example, the workings of the world, of oneself, of others; active pleasures such as creative work and play; and passive pleasures such as appreciating beauty.

From these basic activities we can determine the fundamental conditions for pursuing a good life. The fundamental conditions are the various goods, capacities, and options that human beings qua human beings need whatever else they qua individuals might need in order to pursue the basic activities. For example, the fundamental goods are resources that human beings qua human beings need in order to sustain themselves corporeally, such as food, water, and air. The fundamental capacities are the powers and abilities that human beings qua human beings require whatever else they qua individuals might require in order to pursue the basic activities. These include the capacity to think, to be motivated by facts, to know, to choose an act freely (liberty), to appreciate the worth of something, to develop interpersonal relationships, and to have control of the direction of one's life (autonomy). The fundamental options are the social forms and institutions that human beings qua human beings require in order to exercise their essential capacities to engage in the basic activities. These include the option to have social interaction, to acquire further knowledge, to evaluate and appreciate things, and to determine the direction of one's life.

Having the fundamental conditions for pursuing a good life of course cannot guarantee that an individual has a good life; no condition can guarantee this. Rather, these goods, capacities, and options enable human beings to pursue the basic activities. Also, many of the fundamental conditions are all-purpose conditions in that they are needed for whatever basic activity one aims to pursue. For example, all human beings need food, water, the capacity to think, and the capacity to determine the direction of their lives, whatever basic activity they aim to pursue. In addition, some fundamental conditions may be needed just for pursuing particular basic activities. For instance, the capacity to develop deep personal relationships may be needed only if one aims to pursue deep personal relationships. If so, we can leave it open whether a particular individual will make use of all the fundamental conditions when pursuing a particular kind of good life. Still, an individual should have all the fundamental conditions, because having all the fundamental conditions would enable an individual to pursue any basic activity in a good life. This could become important if, say, an individual changes his or her mind about pursuing a particular kind of good life.[37]

In my view, these fundamental conditions for pursuing a good life ground human rights because having them is of fundamental importance to human beings and because rights can offer powerful protection to those who possess them. The former is true because if anything is of fundamental importance to

human beings, pursuing a characteristically good human life is. It seems clear that if we attach a certain importance to an end, we must attach this importance to the (essential) means to this end, other things being equal. That rights can offer powerful normative protection to those who possess them is well known.[38] By their nature, rights secure the interests of the rightsholders by requiring others, the duty-bearers, to perform certain services for the rightsholders or not to interfere with the rightsholders' pursuit of their essential interests. In addition, at least on certain structural accounts of rights, rights typically prevent the rightsholders' interests that ground rights from being part of a first-order utilitarian calculus. As such, if a rightsholder has a right to something, then typically no non-rights-based considerations can override the rightsholder's right to it.[39] Finally, as some writers have pointed out, because the rightsholders are entitled to certain services in virtue of their rights, rightsholders can simply expect the services without requesting them.[40] Given the strong protection that rights can offer the rightsholders, and given the importance to human beings of having these fundamental conditions, it seems reasonable to think that human beings would have rights to these fundamental conditions. If so, this provides us with an argument for the idea that human beings have human rights to the fundamental conditions for pursuing a good life.

The Fundamental Conditions Approach can explain why many of the rights in the UDHR are genuine human rights. Consider the right to life, liberty, and security of person (Article 3). Whatever else human beings qua individuals need, they qua human beings need life, liberty, and security of person in order to pursue the basic activities. If they are not alive, if they cannot freely choose to act to some degree, or if the security of their person is not guaranteed, they cannot pursue the basic activities. Given this, on the Fundamental Conditions Approach, human beings would have human rights to life, liberty, and security of person.

Consider status rights that protect our moral status as rightsholders such as the right to recognition everywhere as a person before the law (Article 6); the right to equal protection before the law (Article 7); the right against arbitrary arrest, detention, or exile (Article 9); the right to a fair and public hearing (Article 10); and the right to be presumed innocent until proven guilty (Article 11). These are things that human beings qua human beings need whatever else they qua individuals might need in order to pursue the basic activities. In particular, when we pursue the basic activities, conflicts with others are bound to arise. If and when such conflicts arise, we need guarantees that we will be treated fairly and equally. Fair trials, presumption of innocence, equal protection before the law, not being arrested arbitrarily, and so on serve to ensure that we will be treated fairly and equally. As such, they are things that human beings qua human beings need whatever else they qua individuals might need in order to pursue the basic

activities. Given this, the Fundamental Conditions Approach can explain why there are these human rights.

Finally, consider the right to freedom of thought, conscience, and religion (Article 18), the right to freedom of opinion and expression (Article 19), and the right to freedom of peaceful assembly and association (Article 20). As I said earlier, one of the fundamental conditions for pursuing a good life is being able freely to choose to pursue the basic activities. One must have freedom of expression, thought, religion, and association to do so. On the Fundamental Conditions Approach, human beings would have human rights to freedom of thought, religion, expression, and association.[41]

Drawing on the Fundamental Conditions Approach, one might argue that, in an analogous fashion, rightsholding AIs also have rights to the fundamental conditions for pursuing the basic activities. What qualifies as a basic activity for an AI could, in some respects, differ from what qualifies as a basic activity for human beings. For instance, having deep personal relationships may not matter at all to certain kinds of AIs. If so, the fundamental conditions that such an AI would need in order pursue the basic activities would also differ from those of human beings. And supposing that rightsholding AIs are isomorphic, inorganic beings, such AIs would not need food, water, air, or basic medical care in order to pursue the basic activities. This said, rightsholding AIs are still likely to need something functionally equivalent to these fundamental goods, such as fuel and basic maintenance.

In any case, rightsholding AIs are likely to have some of the same rights as human beings. For example, it seems plausible that rightsholding AIs would also need life, liberty, and security of person in order to pursue a good life. On the Fundamental Conditions Approach, rightsholding AIs would have such rights. Likewise, consider status rights such as the right to recognition everywhere as a person before the law; the right to equal protection before the law; the right against arbitrary arrest, detention, or exile; the right to a fair and public hearing; and the right to be presumed innocent until proven guilty. It seems plausible to think that rightsholding AIs would also need these rights in order to pursue the basic activities. Finally, consider the right to freedom of thought and conscience, the right to freedom of opinion and expression, and the right to freedom of peaceful assembly and association. As intelligent thinking beings, it seems that rightsholding AIs would also need these rights in order to pursue the basic activities.

While rightsholding AIs are likely to have some of the same rights as human beings, they could have some of these rights to a different extent. To give an example, consider reproductive rights.[42] With the exception of countries such as China, at least in human history, a (legal) limit has not been placed on human beings with respect to their right to procreate. Though of course for most human

beings, there are natural limits on reproduction, because procreation and child rearing are time- and resource-intensive. These limits might not exist for some rightsholding AIs. For instance, for AIs created through emulations, it could be fairly easy to make many copies of the same emulation whereby the emulated copies would already be "fully grown" AIs. Assuming, though, that running each emulation would use up a significant amount of resources, there might be reasons to impose some limit on a rightsholding AI's right to reproduce. If so, although human beings and rightsholding AIs would both have a right to reproduce, they could have this right to different extents.

Lastly, some rightsholding AIs could have rights that human beings do not have. Consider again a rightsholding AI created through emulation and running on some piece of hardware. As Nick Bostrom and Eliezer Yudkowsky have pointed out, the subjective rate of time for a rightsholding AI could deviate significantly from the rate characteristic of a biological human brain.[43] To understand the concept of subjective rate of time, imagine running an emulated rightsholding AI on a machine that is faster than the one on which it had been running, that is, on an overclocked machine. Suppose further that the rightsholding AI can perceive the external world through some kind of video device. Such an AI should perceive the external world as if it had been slowed down compared to what a human being would perceive. For example, sending and receiving emails might seem instantaneous to me, but for a rightsholding AI running on a significantly overclocked machine, it might seem like eternity. Given that the subjective rate of time for a rightsholding AI could deviate significantly from the rate characteristic of a biological human brain, this could raise the question of whether a rightsholding AI has the right to control its subjective rate of time so that, for example, it could run on the rate of human-biological time. Indeed it could be difficult for a rightsholding AI to participate in social activities and relationships with humans if its subjective rate of time was much faster than the rate for humans. On the other hand, a rightsholding AI might also wish to be able to run at the same (fast) rate as other emulations when it interacts with them. Given the importance of the subjective rate of time for an AI, a case could be made that an AI should have a right to control its subjective rate of time, a right that human beings (at least currently) do not have.

17.5. AI and Greater than Human-Level Moral Status

If AIs can have human-level moral status, can some AIs have greater than human-level moral status? Certainly a number of people believe that AIs could have greater than human-level intelligence. For instance, I. J. Good famously argued that an AI that is intelligent enough to understand its own design could

create a more intelligent successor system, which could then create an even more intelligent successor system, and so on, leading to an "intelligence explosion" or a technological singularity.[44] Could there be an analogous "moral status explosion"? As far as I can see, there does not seem to be any reason why the form of moral status that human beings have, that is, rightsholding, must be the highest form of moral status. This means that at least in theory, some AIs could have higher forms of moral status. Suppose that this is possible. What kind of empirical, nonspeciesist, intrinsic property would ground a form of moral status that is higher than rightsholding? I shall not attempt to answer this question here, but whatever this property may be, it is unlikely that having more of existing intrinsic properties such as intelligence or moral agency gives one greater moral status.

To see this, consider intelligence. Suppose that there are two human beings, Bright and Average. Bright has above-average intelligence, 150 IQ, while Average has average intelligence, 100 IQ. Although Bright has greater intelligence than Average, it seems that they still have the same moral status; that is, they are both rightsholders. Suppose that both Bright and Average are drowning and only one can be saved. Other things being equal, we would not necessarily opt to save one over the other. Importantly, suppose that someone decided to save Average. Other things being equal, we would not think that this individual did something morally impermissible. If so, this suggests that Bright and Average have the same moral status, even though Bright has greater intelligence.

What if there is someone who is much smarter than Average? Call that individual Exceptionally Bright. And let us suppose that she has 350 IQ, while, as before, Average has 100 IQ. Again, it seems that we would think that Exceptionally Bright and Average are both rightsholders. Suppose that Exceptionally Bright and Average are both drowning. Again, other things being equal, we would not necessarily opt to save one over the other. Moreover, suppose that someone decided to save Average instead of Exceptionally Bright. We would not think that this individual did something morally impermissible. If so, this suggests that Exceptionally Bright and Average also have the same moral status, even though Exceptionally Bright has much greater intelligence.[45]

Suppose that this is correct. We can explain why an AI would not have greater moral status in virtue of having much greater intelligence than an average human being. Suppose that Exceptional AI has exceptional intelligence, just like Exceptionally Bright, that is, it has 350 IQ. As before, Average has 100 IQ. Since Exceptionally Bright and Average have the same moral status even though Exceptionally Bright has much greater intelligence than Average, it seems that Exceptional AI would also not have greater moral status than Average just in virtue of having much greater intelligence.

What if an AI has exceptional moral agency instead? As I said earlier, a sufficient condition for human beings being rightsholders is that they have a

physical basis for moral agency. Given this, one might think that having greater moral agency would give an entity greater moral status. However, it does not seem that having greater moral agency gives an entity greater moral status. To see this, suppose that there are two human beings, Joe and Teresa. Joe has average moral agency and Teresa has exceptional moral agency. While Joe is a good friend, keeps his promises, does not lie, cheat, or steal, and so on, Teresa does all this and has devoted her entire life to charity work and the betterment of humanity. Although Teresa has greater moral agency than Joe, it seems that they would still have the same moral status; that is, they are both rightsholders. Now suppose that both Joe and Teresa are drowning and only one can be saved. Other things being equal, we would not necessarily opt to save one over the other. Moreover, suppose that someone decided to save Joe. We would not think that this individual did something morally impermissible. This suggests that Teresa and Joe have the same moral status, even though Teresa has greater moral agency. Similarly, suppose that there is an AI who has exceptional moral agency, just like Teresa. It should also not have greater moral status than Joe. If all of this is right, it remains an open possibility that AIs could have greater than human-level moral status. But just having more of existing intrinsic properties such as intelligence or moral agency will not give an AI greater moral status than that of human beings.

17.6. Summary

As AIs acquire greater capacities, they are likely to acquire greater moral status, raising questions about how we, as moral agents, should treat them. In this chapter I have defended several claims. First, AIs that are alive, conscious, or sentient, or that can feel pain, have desires, and have rational or moral agency should have the same kind of moral status as entities that have the same kind of intrinsic properties. Second, having the physical basis for moral agency is a sufficient condition for an AI to have human-level moral status and be a rightsholder. Third, an AI can come to have the physical basis for moral agency through brain emulation, gradual substitution, or computer algorithms. Fourth, human beings who are interested in becoming AIs and who also care about surviving, numerically speaking, as AIs, should prefer gradual substitution to brain emulation or computer algorithms, as the last two methods would not ensure their survival. Fifth, rightsholding AIs could have rights to the fundamental conditions for pursuing a good life (for an AI of its kind). Some of their rights will be similar to the rights that human beings have in virtue of being human, such as the right to life and liberty and the right to equal protection. They may have also have rights unique to their nature, such as a right to control their subjective rate of time. Sixth, some

AIs could have greater than human-level moral status, but not in virtue of having more of existing intrinsic properties such as intelligence or moral agency.

Much more can be said on this topic of AI and moral status. I hope that what I have said will provide a theoretical framework for thinking about these issues in philosophy and public debates.

Notes

1. For IBM's Watson, see "IBM and 'Jeopardy!' Relive History with Encore Presentation of 'Jeopardy!': The IBM Challenge," *Jeopardy!* website, 2011, https://web.archive.org/web/20130616092431/http://www.jeopardy.com/news/watson1x7ap4.php.
2. For Google DeepMind's AlphaGo, see Richard Lawler, "Google DeepMind AI Wins Final Go Match for 4-1 Series Win," *Engadget*, March 14, 2016, https://www.engadget.com/2016/03/14/the-final-lee-sedol-vs-alphago-match-is-about-to-start/.
3. David Silver et al., "Mastering the Game of Go without Human Knowledge," *Nature* **550** (2017): 354.
4. Samuel Gibbs, "AlphaZero AI Beats Champion Chess Program after Teaching Itself in Four Hours," *Guardian*, December 7, 2017, https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours.
5. Sally Adee, "Say Hello to Machines That Read Your Emotions to Make You Happy," *New Scientist*, May 13, 2015, https://www.newscientist.com/article/mg22630212-900-say-hello-to-machines-that-read-your-emotions-to-make-you-happy/.
6. "Humanity Is Doomed Now That Robots Can Open Doors," video, *RT*, February 12, 2018, https://www.rt.com/news/418600-spotmini-boston-dynamics-robot/.
7. S. Matthew Liao, *The Right to Be Loved* (New York: Oxford University Press, 2015); Frances M. Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007), ch. 7.
8. Andrew Griffin, "Saudi Arabia Grants Citizenship to a Robot for the First Time Ever," *Independent*, October 26, 2017, http://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html.
9. Alex Hern, "Give Robots 'Personhood' Status, EU Committee Argues," *Guardian*, January 12, 2017, https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues.
10. Andrea Morris, "We Need to Talk about Sentient Robots," *Forbes*, March 13, 2018, https://www.forbes.com/sites/andreamorris/2018/03/13/we-need-to-talk-about-sentient-robots/#177156921b2c.
11. This section is partly drawn from Liao, *The Right to Be Loved*.
12. I would like to thank Duncan Purves for this point.
13. For extrinsic, relational approaches to moral status, especially the moral status of AIs, see, e.g., David Gunkel, *The Machine Question* (Cambridge, MA: MIT Press,

2012); Mark Coeckelbergh, *Growing Moral Relations* (Basingstoke, UK: Palgrave Macmillan, 2012).

14. See also S. Matthew Liao, "The Basis of Human Moral Status," *Journal of Moral Philosophy* 7 (2010): 159–79; S. Matthew Liao, "The Genetic Account of Moral Status: A Defense," *Journal of Moral Philosophy* **9** (2012): 265–77.
15. Liao, "The Basis of Human Moral Status."
16. US Genome Project, "Genetic Issues in Mental Retardation: A Report on the Arc's Human Genome Education Project, 1:1," https://web.archive.org/web/20070106152639/http://www.thearc.org/pdf/gbr01.pdf.
17. There are other kinds of genetic defects that involve more than a single gene. As far as I know, either those defects are so severe that the fetuses typically die before birth, or individuals with these defects typically have only mild mental retardation. See, e.g., American College of Obstetricians and Gynecologists, "Prenatal Diagnosis of Fetal Chromosomal Abnormalities," *ACOG Practice Bulletin* 27 (May 2001).
18. Some might suggest that perhaps infants are given preference because they are smaller. However, this cannot be the explanation, since if someone were to have a small turtle, and assuming that turtles are rightsholders, we would not give the turtle preference over an adult human male just because it is smaller.
19. Anders Sandberg, "Feasibility of Whole Brain Emulation," in *Philosophy and Theory of Artificial Intelligence*, ed. Vincent C. Müller (Berlin: Springer, 2013); Anders Sandberg, "Ethics of Brain Emulations," *Journal of Experimental & Theoretical Artificial Intelligence* 26 (2014): 439–57. See also David J. Chalmers, "Uploading: A Philosophical Analysis," in *Intelligence Unbound*, ed. Russell Blackford and Damien Broderick, 102-118 (John Wiley & Sons, 2014).
20. Sandberg. "Ethics of Brain Emulations."
21. See the website of the Human Brain Project: https://www.humanbrainproject.eu/en/.
22. See the website of the Brain Initiative: https://www.braininitiative.nih.gov/?AspxAutoDetectCookieSupport=1.
23. Antonio Regalado, "A Startup Is Pitching a Mind-Uploading Service That Is '100 Percent Fatal,'" *MIT Technology Review*, March 13, 2018, https://www.technologyreview.com/s/610456/a-startup-is-pitching-a-mind-uploading-service-that-is-100-percent-fatal/.
24. S. Matthew Liao, "Twinning, Inorganic Replacement, and the Organism View," *Ratio* **23** (2010): 59–72.
25. See Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2010) for a discussion of these approaches.
26. Ibid.
27. See, e.g., Alison Hills, "Understanding Why," *Nous* **50** (2016): 661–88 for an account of what is involved in understanding why generally and in moral understanding in particular.
28. Noam Chomsky, *Syntactic Structures* (The Hague: Mouton, 1957).
29. John Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1971); John Mikhail, "Universal Moral Grammar: Theory, Evidence and the Future," *Trends*

in Cognitive Sciences **11** (2007): 143–52; John Mikhail, *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment* (New York: Cambridge University Press, 2011).

30. Mikhail, *Elements of Moral Cognition*.
31. Andrew J. Hawkins, "Elon Musk Thinks Humans Need to Become Cyborgs or Risk Irrelevance," *Verge*, February 13, 2017, https://www.theverge.com/2017/2/13/14597434/elon-musk-human-machine-symbiosis-self-driving-cars.
32. See, e.g., S. Matthew Liao, "The Organism View Defended," *Monist* **89** (2006): 334–50.
33. Ibid.
34. Liao, "Twinning, Inorganic Replacement, and the Organism View."
35. Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William M. Ramsey, 316-334 (Cambridge: Cambridge University Press, 2014); Eric Schwitzgebel and Mara Garza, "A Defense of the Rights of Artificial Intelligences," *Midwest Studies in Philosophy* **39** (2015): 98–119.
36. S. Matthew Liao, "Human Rights as Fundamental Conditions for a Good Life," in *Philosophical Foundations of Human Rights*, ed. Rowan Cruft, S. Matthew Liao, and Massimo Renzo, 79-100 (Oxford: Oxford University Press, 2015). This approach assumes that human rights are those that we have simply in virtue of being human and therefore belong to what might be called a Naturalistic Conception of human rights. In recent years a new and purportedly alternative conception of human rights, the so-called Political Conception of human rights, has become increasingly popular. According to the Political Conception, the distinctive nature of human rights is to be understood in light of their role or function in modern international political practice. Elsewhere I have argued that the theoretical distance between the Naturalistic Conception and the Political Conception is not as great as it has been made out to be. See S. Matthew Liao and Adam Etinson, "Political and Naturalistic Conceptions of Human Rights: A False Polemic?," *Journal of Moral Philosophy* **9** (2012): 327–52.
37. In other words, while I am a disjunctivist about basic activities, that is, I do not think that a person has to pursue all the basic activities in order to have a good life, I am a conjunctivist about fundamental conditions, that is, I think that a person has a right to all the fundamental conditions. Also, my notion of fundamental conditions might prompt some to think of Martha Nussbaum's Central Capabilities Approach. Elsewhere I have explained in greater detail how the two views differ. See Liao. "Human Rights as Fundamental Conditions for a Good Life." All too briefly, the hallmark of Nussbaum's approach is her emphasis on our opportunities to choose to do certain things, that is, capabilities, rather than on what we actually choose to do, that is, functionings. However, many human rights cannot be adequately explained in terms of capabilities. For example, in the UDHR, there are a number of human rights that protect our moral status as persons, that is, status rights, such as the right to recognition everywhere as a person before the law (Article 6); the right to equal protection before the law (Article 7); the right against arbitrary arrest, detention, or exile (Article 9); the right to a fair and public hearing (Article 10); the right to be presumed

innocent until proven guilty (Article 11). Nussbaum's approach seems to imply that one can sometimes choose not to exercise these rights, since capabilities are concerned with our real opportunities to choose. But it does not seem that one can sometimes choose whether or not to exercise these rights. For instance, it does not seem that one can sometimes choose not to be recognized everywhere as a person before the law; choose not to have equal protection before the law; choose to be arrested arbitrarily; choose to have an unfair hearing; and choose to be presumed guilty. Hence capabilities do not seem particularly well-suited to explain these rights. In contrast, my approach can explain status rights. When we pursue the basic activities, conflicts with others are bound to arise. If and when such conflicts arise, we need guarantees that we will be treated fairly and equally. A fair trial, presumption of innocence, equal protection before the law, not being arrested arbitrarily, and so on serve to ensure that we will be treated fairly and equally. As such, they are things that human beings qua human beings need whatever else they qua individuals might need in order to pursue the basic activities. As such, the approach I advocate can explain why there are these human rights.

38. Rights could also have noninstrumental importance in addition to having instrumental importance.
39. Ronald Dworkin, *Taking Rights Seriously* (London: Duckworth, 1977).
40. Joel Feinberg, "The Nature and Value of Rights," in *Bioethics and Human Rights: A Reader for Health Professionals*, ed. Elsie Bandman and Bertram Bandman (Boston: Little, Brown, 1970).
41. It may be worth mentioning that the Fundamental Conditions Approach would exclude some of the claims in the UDHR as genuine human rights. See S. Matthew Liao, "A Short Introduction to the Ethics of Artificial Intelligence" (this volume), for a discussion of why there is not a right to periodic holidays with pay (Article 24).
42. Bostrom and Yudkowsky, "The Ethics of Artificial Intelligence."
43. Ibid.
44. Irving John Good, "Speculations concerning the First Ultraintelligent Machine," in *Advances in Computers*, ed. Alt and Rubinoff (New York: Academic Press, 1965).
45. I would be inclined to say the same thing even if there was an individual with an IQ of 1000.

References

Adee, Sally. "Say Hello to Machines That Read Your Emotions to Make You Happy." *New Scientist*, May 13, 2015. https://www.newscientist.com/article/mg22630212-900-say-hello-to-machines-that-read-your-emotions-to-make-you-happy/.

American College of Obstetricians and Gynecologists. "Prenatal Diagnosis of Fetal Chromosomal Abnormalities." *ACOG Practice Bulletin* 27 (May 2001).

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey, 316-334. Cambridge: Cambridge University Press, 2014.

Chalmers, David J. "Uploading: A Philosophical Analysis." In *Intelligence Unbound*, edited by Russell Blackford and Damien Broderick, 102-118. John Wiley & Sons, 2014.

Chomsky, Noam. *Syntactic Structures*. The Hague: Mouton, 1957.

Coeckelbergh, Mark. *Growing Moral Relations*. Basingstoke, UK: Palgrave Macmillan, 2012.

Dworkin, Ronald. *Taking Rights Seriously*. London: Duckworth, 1977.

Feinberg, Joel. "The Nature and Value of Rights." In *Bioethics and Human Rights: A Reader for Health Professionals*, edited by Elsie Bandman and Bertram Bandman. Boston: Little, Brown, 1970.

Gibbs, Samuel. "AlphaZero AI Beats Champion Chess Program after Teaching Itself in Four Hours." *Guardian*, December 7, 2017. https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours.

Good, Irving John. "Speculations concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Alt and Rubinoff. New York: Academic Press, 1965.

Griffin, Andrew. "Saudi Arabia Grants Citizenship to a Robot for the First Time Ever." *Independent*, October 26, 2017. http://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html.

Gunkel, David. *The Machine Question*. Cambridge, MA: MIT Press, 2012.

Hawkins, Andrew J. "Elon Musk Thinks Humans Need to Become Cyborgs or Risk Irrelevance." *Verge*, February 13, 2017. https://www.theverge.com/2017/2/13/14597434/elon-musk-human-machine-symbiosis-self-driving-cars.

Hern, Alex. "Give Robots 'Personhood' Status, EU Committee Argues." *Guardian*, January 12, 2017. https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues.

Hills, Alison. "Understanding Why." *Nous* 50 (2016): 661–88.

"Humanity Is Doomed Now That Robots Can Open Doors." Video. *RT*, February 12, 2018. https://www.rt.com/news/418600-spotmini-boston-dynamics-robot/.

"IBM and 'Jeopardy!' Relive History with Encore Presentation of 'Jeopardy!': The IBM Challenge." *Jeopardy!* website, 2011. https://web.archive.org/web/20130616092431/http://www.jeopardy.com/news/watson1x7ap4.php.

Kamm, Frances M. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.

Liao, S. Matthew. "The Basis of Human Moral Status." *Journal of Moral Philosophy* 7 (2010): 159–79.

Liao, S. Matthew. "The Genetic Account of Moral Status: A Defense." *Journal of Moral Philosophy* 9 (2012): 265–77.

Liao, S. Matthew. "Human Rights as Fundamental Conditions for a Good Life." In *Philosophical Foundations of Human Rights*, edited by Rowan Cruft, S. Matthew Liao, and Massimo Renzo, 79-100. Oxford: Oxford University Press, 2015.

Liao, S. Matthew. "The Organism View Defended." *Monist* 89 (2006): 334–50.

Liao, S. Matthew. *The Right to Be Loved*. New York: Oxford University Press, 2015.

Liao, S. Matthew. "Twinning, Inorganic Replacement, and the Organism View." *Ratio* 23 (2010): 59–72.

Liao, S. Matthew, and Adam Etinson. "Political and Naturalistic Conceptions of Human Rights: A False Polemic?" *Journal of Moral Philosophy* 9 (2012): 327–52.

Lawler, Richard. "Google DeepMind AI Wins Final Go Match for 4-1 Series Win." *Engadget*, March 14, 2016. https://www.engadget.com/2016/03/14/the-final-lee-sedol-vs-alphago-match-is-about-to-start/.

Mikhail, John. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. New York: Cambridge University Press, 2011.

Mikhail, John. "Universal Moral Grammar: Theory, Evidence and the Future." *Trends in Cognitive Sciences* 11 (2007): 143–52.

Morris, Andrea. "We Need to Talk about Sentient Robots." *Forbes*, March 13, 2018. https://www.forbes.com/sites/andreamorris/2018/03/13/we-need-to-talk-about-sentient-robots/#177156921b2c.

Rawls, John. *A Theory of Justice*. Oxford: Oxford University Press, 1971.

Regalado, Antonio. "A Startup Is Pitching a Mind-Uploading Service That Is '100 Percent Fatal.'" *MIT Technology Review*, March 13, 2018. https://www.technologyreview.com/s/610456/a-startup-is-pitching-a-mind-uploading-service-that-is-100-percent-fatal/.

Sandberg, Anders. "Ethics of Brain Emulations." *Journal of Experimental & Theoretical Artificial Intelligence* 26 (2014): 439–57.

Sandberg, Anders. "Feasibility of Whole Brain Emulation." In *Philosophy and Theory of Artificial Intelligence*, edited by Vincent C. Müller. Berlin: Springer, 2013.

Schwitzgebel, Eric, and Mara Garza. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39 (2015): 98–119.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go without Human Knowledge." *Nature* 550 (2017): 354.

US Genome Project. "Genetic Issues in Mental Retardation: A Report on the Arc's Human Genome Education Project, 1:1." https://web.archive.org/web/20070106152639/http://www.thearc.org/pdf/gbr01.pdf.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2010.

Index

Tables and figures are indicated by *t* and *f* following the page number

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.