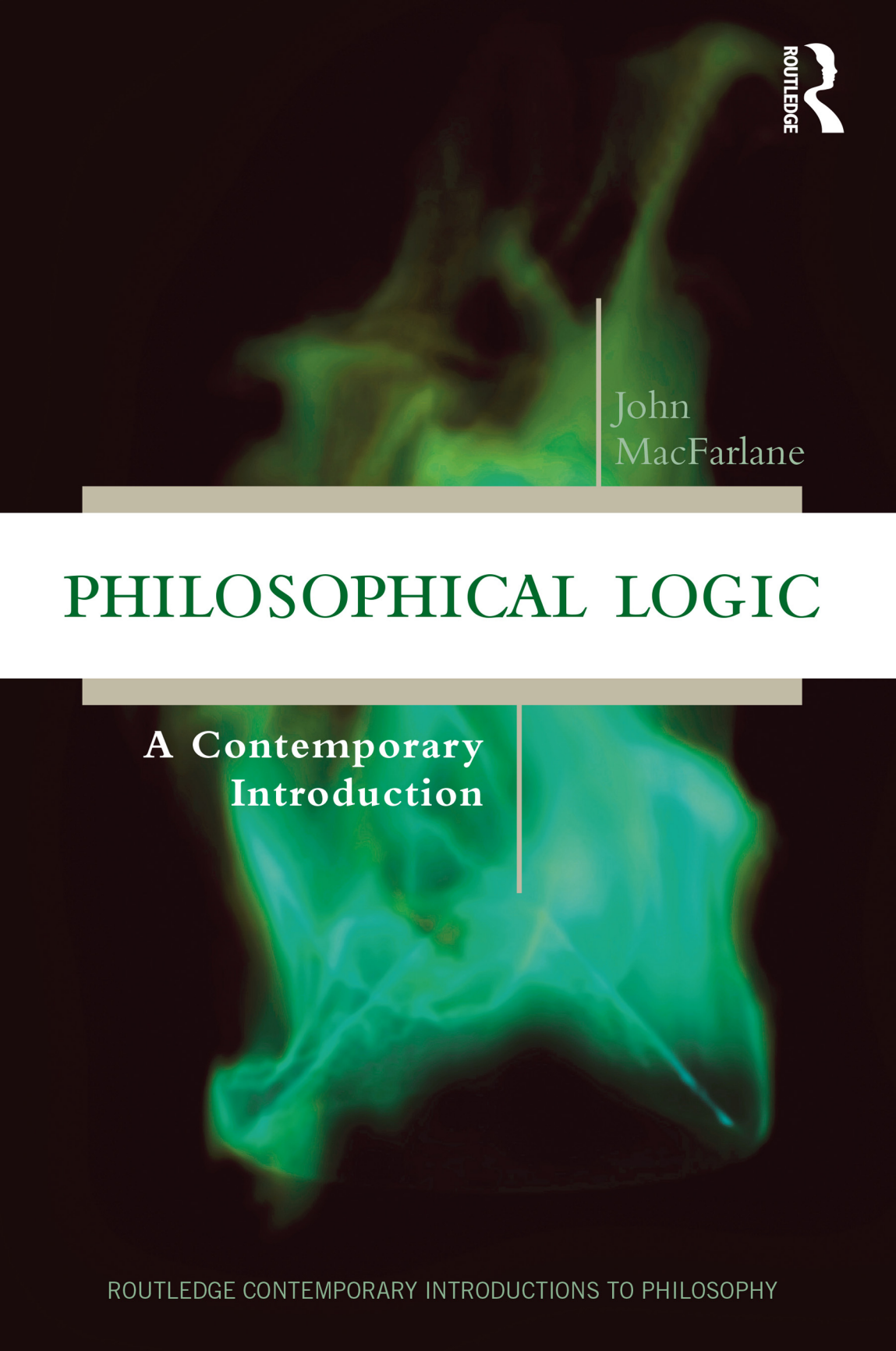

ROUTLEDGE
John
MacFarlane
PHILOSOPHICAL LOGIC
A Contemporary
Introduction
ROUTLEDGE CONTEMPORARY INTRODUCTIONS TO PHILOSOPHY

"This is the perfect book for coverage of classic debates in mainstream philosophy of logic. It's also the perfect source for exceptionally clear reviews of standard logical machinery (e.g., standard modal machinery, quantifier machinery, higher-order machinery, etc.). Very user-friendly, clear, and accurate on all of the topics that it covers, this is my new required text for classic debates in the philosophy of logic."

Jc Beall, *University of Notre Dame*

"John MacFarlane displays his usual lively and engaging writing style, and is neutral on controversial issues, giving the arguments employed by both sides. It is an excellent overview of some key topics in the field."

Stewart Shapiro, *Ohio State University*

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Philosophical Logic

Introductory logic is generally taught as a straightforward technical discipline. In this book, John MacFarlane helps the reader think about the limitations of, presuppositions of, and alternatives to classical first-order predicate logic, making this an ideal introduction to philosophical logic for any student who already has completed an introductory logic course.

The book explores the following questions. Are there quantificational idioms that cannot be expressed with the familiar universal and existential quantifiers? How can logic be extended to capture modal notions like necessity and obligation? Does the material conditional adequately capture the meaning of 'if'—and if not, what are the alternatives? Should logical consequence be understood in terms of models or in terms of proofs? Can one intelligibly question the validity of basic logical principles like Modus Ponens or Double Negation Elimination? Is the fact that classical logic validates the inference from a contradiction to anything a flaw, and if so, how can logic be modified to repair it? How, exactly, is logic related to reasoning? Must classical logic be revised in order to be applied to vague language, and if so how? Each chapter is organized around suggested readings and includes exercises designed to deepen the reader's understanding.

Key Features:

- An integrated treatment of the technical and philosophical issues comprising philosophical logic
- Designed to serve students taking only one course in logic beyond the introductory level
- Provides tools and concepts necessary to understand work in many areas of analytic philosophy
- Includes exercises, suggested readings, and suggestions for further exploration in each chapter

John MacFarlane is Professor of Philosophy and a member of the Group in Logic and the Methodology of Science at the University of California, Berkeley. He is the author of *Assessment Sensitivity: Relative Truth and Its Applications* (2014).

ROUTLEDGE CONTEMPORARY INTRODUCTIONS TO PHILOSOPHY

Series editor:
Paul K. Moser
Loyola University of Chicago

This innovative, well-structured series is for students who have already done an introductory course in philosophy. Each book introduces a core general subject in contemporary philosophy and offers students an accessible but substantial transition from introductory to higher-level college work in that subject. The series is accessible to non-specialists and each book clearly motivates and expounds the problems and positions introduced. An orientating chapter briefly introduces its topic and reminds readers of any crucial material they need to have retained from a typical introductory course. Considerable attention is given to explaining the central philosophical problems of a subject and the main competing solutions and arguments for those solutions. The primary aim is to educate students in the main problems, positions and arguments of contemporary philosophy rather than to convince students of a single position.

Recently Published Volumes:

Philosophy of Language
3rd Edition
William G. Lycan

Philosophy of Mind
4th Edition
John Heil

Philosophy of Science
4th Edition
Alex Rosenberg and Lee McIntyre

Philosophy of Western Music
Andrew Kania

Phenomenology
Walter Hopp

Philosophical Logic
John MacFarlane

For a full list of published Routledge Contemporary Introductions to Philosophy, please visit https://www.routledge.com/Routledge-Contemporary-Introductions-to-Philosophy/book-series/SE0111

Philosophical Logic

A Contemporary Introduction

John MacFarlane

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2021
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this title has been requested

ISBN: 978-1-138-73764-8 (hbk)
ISBN: 978-1-138-73765-5 (pbk)
ISBN: 978-1-315-18524-8 (ebk)

Typeset in EB Garamond, Garamond-Math, and Gill Sans by the author.

Publisher's Note
This book has been prepared from camera-ready copy provided by the author.

Contents

List of Exercises

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

If you tried to figure out what philosophical logic was by looking in the literature, you might easily become confused. John Burgess characterizes philosophical logic as a branch of formal logic: "Philosophical logic as understood here is the part of logic dealing with what classical logic leaves out, or allegedly gets wrong" (Burgess 2009, p. 1). Sybil Wolfram, by contrast, sets philosophical logic apart from formal logic: "Rather than setting out to codify valid arguments and to supply axioms and notations allowing the assessment of increasingly complex arguments, it examines the bricks and mortar from which such systems are built"—for example, meaning, truth, and proposition (Wolfram 1989, pp. 2–3). Their textbooks, both entitled *Philosophical Logic*, cover entirely different subject matters.

The root of this confusion is that the term 'philosophical logic' is ambiguous. Just as 'mathematical logic' means both (a) the mathematical investigation of basic notions of logic and (b) the deployment of logic to help with mathematical problems, so 'philosophical logic' means both (a) the philosophical investigation of the basic notions of logic and (b) the deployment of logic to help with philosophical problems. In the first sense, philosophical logic is the philosophy of logic: the investigation of the fundamental concepts of logic. In the second sense, it consists largely in the formal investigation of alternatives and extensions to classical logic.

It is common to avoid the ambiguity, as Wolfram and Burgess both do, by using 'philosophical logic' in one sense or the other. But in this text we embrace the ambiguity, introducing students to philosophical logic in both its senses. On the one hand, students will consider philosophical questions about truth values, logical consequence, *de re* modality, fundamental logical disagreement, and the relation of logic to reasoning. And on the other hand, they will learn about modal logic, intuitionistic logic, relevance logic, plural and substitutional quantifiers, conditionals, and vagueness.

Why approach things this way, rather than focusing on one side of the ambiguity? Because doing each well requires doing the other. For example, relevance logic and intuitionistic logic are best motivated by reflection on the notion of logical consequence and the way it is explicated in classical logic. The assessment of these logics, too, depends on philosophical issues about logical consequence. So it is

quite artificial to separate the study of nonclassical logics from the philosophical study of the basic notions of logic.

On the other hand, many of the philosophical questions about the basic building blocks of logic can only be properly discussed once we have some nonclassical logics clearly in view. For example, thinking clearly about whether relevance should be required for logical consequence requires understanding the tradeoffs one would need to make in actually developing a relevance logic. Any discussion of the meaning of truth values requires us to see the role truth values might play in a multi-valued logic. And discussions of modality and propositions can be illuminated by a close examination of the slingshot argument, which requires a bit of instruction in modal logic and quantification.

This book is meant for advanced undergraduates or graduate students who have taken a first course in symbolic logic. It aims to impart a sense of the limits of first-order logic, a familiarity with and facility with logical systems, and an understanding of some of the important philosophical issues that can be raised about logic. A side benefit will be increased comprehension of work in other areas of analytic philosophy, in which a certain amount of "logical culture" is often taken for granted.

The chapters of this book have been arranged in what seems to me the most sensible order, but it should be possible to plot various courses through the material, as the chapters are only lightly coupled. Each chapter is built around some readings, which should be read in conjunction with the chapter. (At the end of the chapter there are some suggestions for further reading, for students who want to go deeper.) Each chapter also contains exercises, which are designed to help students think more deeply about the material. (Exercises marked with a $\star$ are harder and more open-ended; in a course they might be made optional.)

Chapter 1: Fundamentals. We presuppose that readers will have had an introductory course in symbolic logic. But since introductory courses vary greatly both in what they cover and in how they cover it, we begin with a brief review of propositional logic and first-order predicate logic with identity, covering syntax, semantics, natural deduction proofs, and definitions of basic concepts.

Chapter 2: Quantifiers. Students often have the impression that the existential and universal quantifiers they learned in introductory logic suffice for the formalization of all quantificational idioms. We will see that this isn't the case and explore some ways the machinery of quantification might be extended or reinterpreted. Topics include definite descriptions, generalized quantifiers, second-order quantifiers, and substitutional quantifiers.

Chapter 3: Modal Logic. In addition to talking about what *is* the case, we talk about what might have been the case and what could not have been otherwise. Modal logic gives us tools to analyze reasoning involving these notions. We will acquire a basic grasp of the fundamentals of propositional modal logic (syntax, semantics, and proofs), and look at some different ways the modalities might be interpreted. We will then delve into some hairy conceptual problems surrounding *quantified* modal logic, explored by W. V. O. Quine, Saul Kripke, and others. We will also look at the famous *slingshot argument*, which was used by Quine and Donald Davidson to reject modal logic and correspondence theories of truth. (Assessing this argument will require bringing together our work on modal logic with our work on quantifiers.)

Chapter 4: Conditionals. In introductory logic classes one is taught to translate English conditionals using the material conditional, a truth-functional connective. This leads to some odd results: for example, it implies that 'If the coin landed heads, Sam won ten dollars' is true if the coin landed tails—even if Sam only bet one dollar on heads. We will consider some attempts to defend the material-conditional analysis of indicative conditionals in English. Then we will consider two alternatives: Dorothy Edgington's view that indicative conditionals have no truth-conditions and Robert Stalnaker's influential modal account. Finally, we will look at Vann McGee's "counterexample to Modus Ponens," and consider whether this sacrosanct inference rule is actually invalid.

Chapter 5: Logical Consequence via Models. Logic is sometimes described as the study of what follows from what—that is, of logical consequence. But how should we think of this relation? Different ways of explicating consequence seem to have different implications for what follows from what, and hence different implications for formal logic. In this chapter we will look at Alfred Tarski's account of logical consequence, which has become the orthodox account. On this account, logical consequence is a matter of *truth preservation*: P follows from Q if there is no model in which P is true and Q false. We will discuss how this account relates to the older idea that P follows from Q if it is *impossible* for P to be true and Q false, and how it makes consequence relative to a choice of *logical constants*. We will also consider some criticisms of this account.

Chapter 6: Logical Consequence via Proofs. Instead of thinking of the meaning of logical constants semantically—for example, as truth functions—we might try to understand them in terms of the stipulated rules governing their use. The logical consequences of a set of premises can then be defined as the sentences that can be proven from them using only the rules that define the logical constants. We

consider some classic objections to this strategy, and look at how Dag Prawitz overcomes them in his proof-theoretic account of logical consequence. Prawitz's account yields a nonclassical logic, *intuitionistic logic*. The dispute between classical and intuitionistic logicians about basic inference forms like Double Negation Elimination is a paradigm example of fundamental logical disagreement. We will consider to what extent this disagreement can be thought of as a verbal one, about the meanings of the logical connectives.

Chapter 7: Relevance, Logic, and Reasoning. Students often find it counterintuitive that, in classical logic, anything follows from a contradiction. One source of resistance is the idea that the premises of a valid argument must be *relevant* to the conclusion. We will look at several ways to develop nonclassical logics that respect this idea, with the aim of getting a sense of the costs of imposing a requirement of relevance. Then we will look more carefully at the motivation for relevance logic, with attention to how logic relates to reasoning, and to how a logic that allows statements to be *both* true and false might be interpreted.

Chapter 8: Vagueness and the Sorites Paradox. The ancient *sorites* paradox, or paradox of the heap, concludes that one grain of sand makes a heap, since 5000 grains of sand make a heap, and taking a single grain of sand from a heap cannot make it a non-heap. Some philosophical logicians have suggested that it is a mistake to use classical logic and semantics in analyzing this argument, and they have proposed a number of alternatives. We will consider three of them: (a) a three-valued logic, (b) a continuum-valued (or fuzzy) logic, and (c) a supervaluational approach that preserves classical logic but not classical semantics. We will also look at a short argument by Gareth Evans that purports to show that vagueness must be a semantic phenomenon: that is, that there is no vagueness "in the world."

Acknowledgements

This book had its genesis in a Philosophical Logic course I have been teaching at Berkeley, on and off, for more than a decade. I am grateful to all of the students who have taken this course for helping me see what works and what doesn't. I am also grateful to Kenny Easwaran, Fabrizio Cariani, Justin Bledin, Justin Vlasits, and James Walsh, who all served as teaching assistants for the course and gave me much helpful feedback.

For invaluable comments on the entire manuscript I am grateful to James Walsh and an anonymous reviewer for Routledge. Wesley Holliday gave me useful feedback on the proof system in Chapter 1. Andy Beck, my editor at Routledge, deserves credit for proposing the project in the first place and helping it to completion. To typeset the book I relied on LaTeX, and in particular the excellent `memoir` package.

I began the book while on sabbatical in Paris in the 2016/17 academic year. I am very grateful for a fellowship from the Paris Institute for Advanced Studies, with the financial support of the French State managed by the Agence Nationale de la Recherche, programme "Investissements d'avenir," (ANR-11-LABX-0027-01 Labex RFIEA+), and the Fondation Maison des Sciences de l'Homme. I am also grateful to UC Berkeley for a Humanities Research Fellowship.

My most basic debt is to Nuel Belnap. Without his brilliant logic pedagogy I would never have gotten interested in philosophical logic.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1 Fundamentals

In general, the task of describing a logical system comes in three parts:

Grammar Describing what counts as a formula.

Semantics Defining truth in a model (and, derivatively, logical consequence and related notions).

Proofs Describing what counts as a proof.

In this chapter, we will go through these three parts for propositional logic, predicate logic, and the logic of identity. We will also review the distinction between use and mention and introduce Quine's device of "quasiquotation," which we will need later to keep from getting confused.

We assume you have taken a first course in symbolic logic, covering propositional and predicate logic (but not metalogical results like soundness and completeness). But, because such courses differ considerably in the symbols, terminology, and proof system they use, a brief review of these fundamentals will help ensure that everyone is on the same page with the basics.

Some or all of this may be old hat. Other things may be unfamiliar. Many logic textbooks do not give a rigorous account of the semantics of first-order logic, and many do not teach the Fitch-style natural deduction proofs used here. Sometimes courses in predicate logic do not cover identity at all. Before going further in this book, you should be comfortable doing exercises of the sort given in this chapter.

1.1 Propositional logic

1.1.1 Grammar

- A *propositional constant* (a capital letter, possibly with a numerical subscript) is a formula. There are infinitely many propositional constants: A, B_{15}, Z_{731}, etc.
- $\bot$ is a formula.

- If p and q are formulas, then $(p \vee q)$, $(p \wedge q)$, $(p \supset q)$, $(p \equiv q)$, and $\neg p$ are formulas.
- Nothing else is a formula.

You might have used different symbols in your logic class: & or • for conjunction, ~ or – for negation, $\rightarrow$ for the conditional, $\leftrightarrow$ for the biconditional. You might not have seen $\perp$ (called *bottom*, *falsum*, or *das Absurde*): we will explain its meaning shortly.

The lowercase letters p and q are not propositional constants. They are used to mark places where an arbitrary formula may be inserted into a *schema*: a pattern that many different formulas can "fit" or "instantiate." Here are some instances of the schema $(p \wedge (p \vee q))$:

(1) $(\overbrace{(A \wedge B)}^{p} \wedge (\overbrace{(A \wedge B)}^{p} \vee \overbrace{\neg A}^{q}))$

(2) $(\overbrace{(A \vee \neg B)}^{p} \wedge (\overbrace{(A \vee \neg B)}^{p} \vee \overbrace{(A \vee \neg B)}^{q}))$

In (1), we substituted the formula $(A \wedge B)$ for the letter p in the schema, and we substituted the formula $\neg A$ for q. In (2), we substituted $(A \vee \neg B)$ for both p and q. Can you see why the following formulas are *not* instances of $(p \wedge (p \vee q))$?

(3) $(A \wedge (B \vee C))$

(4) $(\neg A \vee (\neg A \wedge B))$

Convention for parentheses. The parentheses can get a bit bothersome, so we will adopt the following conventions:

- Outer parentheses may be dropped: so, for example, $A \vee B$ is an abbreviation for $(A \vee B)$.
- We will consider $(p \vee q \vee r)$ as an abbreviation for $((p \vee q) \vee r)$, and $(p \wedge q \wedge r)$ as an abbreviation for $((p \wedge q) \wedge r)$

1.1.2 Semantics

Logicians don't normally concern themselves much with *truth* simpliciter. Instead, they use a relativized notion of truth: *truth in a model*. You may not be familiar with this terminology, but you should be acquainted with the idea of *truth in a row of a truth table*, and in (classical) propositional logic, that is basically what truth in a model amounts to.

A *model* is something that provides enough information to determine truth values for *all* of the formulas in a language. How much information is required

depends on the language. In the simple propositional language we're considering, we have a very limited vocabulary—propositional constants and a few truth-functional connectives—and that allows us to use very simple models. When we add quantifiers, and, later, modal operators, we will need more complex models.

What does it mean to say that a connective is *truth-functional*? It means that the only information we need to determine the truth value of a compound formula formed with one of these connectives is the truth values of the formulas it connects. Thus, for example, all we need to know to determine the truth value of $\neg B \wedge C$ are the truth values of $\neg B$ and C. And all we need to know to determine the truth value of $\neg B$ is the truth value of B. No further information about the meaning of B is needed.

Because all of our connectives are truth functional, once the truth values of the propositional constants are fixed, the truth values of *all* the formulas in the language are fixed as a result. Because of this, a *model* for classical propositional logic can be just an assignment of truth values (true or false) to each propositional constant.

Although there are infinitely many propositional constants, usually we only need to concern ourselves with a few of them—those that occur in the arguments we're analyzing. Suppose the formulas we're looking at contain the constants A, B, and C. Then we can describe two different models (v_1 and v_2) by describing the truth values they give to these constants:

$$v_1(A) = \text{True}, v_1(B) = \text{False}, v_1(C) = \text{False}$$
$$v_2(A) = \text{False}, v_2(B) = \text{False}, v_2(C) = \text{True}$$

This notation is a bit tedious, though. We can present the same information in tabular form:

	A	B	C
v_1	T	F	F
v_2	F	F	T

You can see that a model is basically a row of a truth table.[1]

Why are logicians interested in *truth in a model*? Because all of the fundamental semantic logical relations are defined in terms of it:

[1] "Basically," because in fact a row of a truth table represents infinitely many models that agree on their assignments to the propositional constants represented in the table, but disagree on their assignments to propositional constants not listed. We can safely ignore this subtlety for most purposes, because assignments to propositional constants not contained in a formula are irrelevant to its truth in a model.

An argument[2] is *valid* iff there is no model in which all of its premises are true and its conclusion false. In this case the conclusion is said to be a *logical consequence* of the premises.

A formula p *implies* another formula q iff there is no model in which p is true and q is false.

Two formulas are *equivalent* iff they have the same truth value in every model.

A set of formulas is *satisfiable* iff there is a model in which all are true.

A formula p is a *logical truth* if it is true in every model, a *logical contradiction* or *logical falsehood* if it is false in every model, and *logically contingent* if it is neither a logical truth nor a contradiction.

Sometimes the terms defined above are qualified to indicate the *kind* of models we are considering. For example, when we are considering only models of classical propositional logic, where all the connectives are truth-functional, we can talk of "truth-functional validity," "truth-functional equivalence," and so on, to make that clear. The term *tautology* is sometimes used for truth-functional logical truth. As we've seen, in classical propositional logic, a model is just a row of a truth table. So, in classical propositional logic, a tautology is a formula that is true in all rows of a truth table; two formulas are equivalent iff they have the same truth values in each row of a truth table, and so on.

To give the semantics of our language, we need to define truth in a model v for arbitrary formulas:

- When p is a propositional constant, p is true in v iff $v(p) = \text{True}$.
- $\bot$ is not true in any model v.
- $\neg p$ is true in v iff p is not true in v.
- $p \wedge q$ is true in v iff p is true in v and q is true in v.
- $p \vee q$ is true in v iff p is true in v or q is true in v (or both).
- $p \supset q$ is true in v iff p is not true in v or q is true in v.
- $p \equiv q$ is true in v iff either both p and q are true in v or neither p nor q is true in v.

[2] An *argument*, in the logician's sense, is just a pair consisting of a set of premises and a conclusion. This is a departure from the ordinary sense of 'argument', which is usually used either for a dispute or for the reasoning that connects the premises with the conclusion. The logician's notion of *proof* is related to this latter sense.

Exercise 1.1: Basic concepts

1. Give an example of a truth-functional connective other than the usual ones (conjunction, disjunction, negation, material conditional and biconditional). Explain what makes it truth-functional. Give an example of a non-truth-functional connective, and show that it is not truth-functional.

2. Write out truth tables for the following formulas:

   a) $P \vee \neg(R \equiv S)$

   b) $Q \vee \neg(\neg Q \wedge \neg R)$

3. What does it mean to say that a formula of propositional logic is a *tautology*? A *contradiction*? *Contingent*? In which categories do the following formulas fall?

   a) $P \supset (\bot \supset \neg P)$

   b) $P \vee (Q \wedge (\neg P \vee \neg Q))$

   Note: $\bot$ is a special propositional constant (or, if you like, a 0-place connective—a connective that takes 0 formulas and yields a new formula). It is False in every model, so when you do your truth tables, you can just write F in every row under $\bot$.

4. Is the following set of formulas satisfiable?

   $P \supset Q, Q \supset S, \neg S \supset \neg P$

5. What does it mean to say that two formulas are *logically equivalent*? Give an (interesting) example of two logically equivalent formulas of propositional logic.

6. Does $P \supset (Q \wedge \neg Q)$ truth-functionally imply $\neg P \vee R$? Does $P \supset (Q \supset R)$ truth-functionally imply $R \supset (\neg Q \supset \neg P)$?

These clauses, which express the information encoded in the classical truth tables, determine a truth value for any formula built up from propositional constants and $\bot$ using $\wedge$, $\vee$, and $\neg$. For example:

$((A \wedge B) \vee \neg A)$ is true in v

iff $(A \wedge B)$ is true in v or $\neg A$ is true in v (by the clause for $\vee$)

iff (A is true in v and B is true in v) or $\neg A$ is true in v (by the clause for $\wedge$)

iff (A is true in v and B is true in v) or A is not true in v (by the clause for $\neg$)

iff ($v(A) = \text{True}$ and $v(B) = \text{True}$) or it is not the case that $v(A) = \text{True}$ (by the clause for propositional constants).

1.1.3 Proofs

There are many different proof systems for propositional logic. Natural deduction systems try to formalize patterns of ordinary logical reasoning. In your introductory logic course, you might have learned a "Lemmon-style system," in which numbers are used to keep track of the hypotheses on which a given line depends. I favor "Fitch-style systems," which keep track of undischarged assumptions using vertical lines, rather than numbers.[3] This geometrical presentation makes the hypothesis structure of a proof more perspicuous.

Structural Rules

Fitch-style systems have two kinds of rules. First, there are *structural rules*, which concern structural aspects of proofs and do not involve any specific connectives or quantifiers.

Hyp

Any formula may be written down at any time, above a horizontal line. The justification may be written "Hyp" (for "hypothesis"), or the justification may simply be omitted, since it is clear from the horizontal line itself. Alternatively, several formulas may be simultaneously hypothesized, one per line, with a horizontal line below them (see Example 1.1).

[3] These are named after F. B Fitch, who invented them (Fitch 1952). I learned Fitch-style deductions from Fitch's student Nuel Belnap, and my presentation here draws on his unpublished textbook (Belnap 2009) and on Barwise and Etchemendy 1999.

1	$S \wedge T$	Hyp
2	Q	Hyp
3	⋮	
4	⋮	
5	R	Hyp
6	⋮	
7	P	Hyp
8	⋮	
9	⋮	
10	⋮	

(1.1)

Each hypothesis begins a *subproof*, which we signify by a vertical line to the left of the formulas. Subsequent steps in the subproof are considered to be proved "under" the hypothesis (or hypotheses), not proved outright; that is, they are asserted as true under the supposition that the hypothesis is true, not as categorically true. In Example 1.1, a formula at line 3, 4, or 10 is being asserted as true on the assumptions $S \wedge T$ and Q; a formula at line 6 is being asserted as true on the assumptions $S \wedge T$, Q, and R; and a formula at line 8 is being asserted as true on the assumptions $S \wedge T$, Q, R, and P.

Subproofs may be nested. A subproof occurring inside another subproof is said to be *subordinate* to it (and the containing subordinate is *superordinate* to the one contained). In Example 1.1, the subproof that extends from lines 5–9 is subordinate to the subproof that extends from lines 1–10. And the subproof that extends from lines 7–8 is subordinate to both the subproof that extends from lines 5–9 and to the subproof that extends from lines 1–10.

Reit

The "Reit" rule allows you to reiterate any formula into any *subordinate* subproof. Here is an example:

1	$S \wedge T$	Hyp	
2	$\quad R$	Hyp	
3	$\quad S \wedge T$	Reit 1	(1.2)
4	$\quad\quad Q$	Hyp	
5	$\quad\quad S \wedge T$	Reit 1	

The Reit rule is needed because our rules for the connectives (to be given below) require the premises to be in the same subproof. The natural deduction system you learned may not have had a Reit rule: some systems allow rules to use premises in superordinate subproofs without bringing them together into the same subproof through explicit reiteration. In §3.1.4, when we study natural deduction systems for modal logic, we will see the point of keeping track of reiteration explicitly

To avoid tedium, if a formula can be derived using another rule, together with one or more obvious applications of Reit, we will allow these rules to be "collapsed," with the justification mentioning both the other rule and "+ Reit." Thus, instead of

1	P	Hyp		
2	$\quad R$	Hyp		
3	$\quad\quad Q$	Hyp		(1.3)
4	$\quad\quad P$	Reit 1		
5	$\quad\quad P \wedge Q$	$\wedge$ Intro	3, 4	

one can write:

1	P	Hyp	
2	$\quad R$	Hyp	
3	$\quad\quad Q$	Hyp	(1.4)
4	$\quad\quad P \wedge Q$	$\wedge$ Intro + Reit 1, 3	

Rules for propositional connectives

The remaining rules concern specific connectives. In general, we will have two rules for each connective: one to introduce the connective, and one to eliminate it. (These paired rules are sometimes called *intelim rules*.) The exceptions to this generalization are $\bot$, which has no introduction rule, and $\neg$, which has two distinct elimination rules.

$\wedge$ Intro

If a subproof contains a formula p and a formula q, you may write down $p \wedge q$ in the same subproof with the justification "$\wedge$ Intro." Example:

1	P	Hyp
2	$R \vee S$	Hyp
3	$P \wedge (R \vee S)$	$\wedge$ Intro 1, 2

(1.5)

$\wedge$ Elim

If a formula $p \wedge q$ occurs in a subproof, you may write down either p or q in the same subproof with the justification "$\wedge$ Elim." Example:

1	$R \wedge S$	Hyp
2	S	$\wedge$ Elim 1

(1.6)

$\supset$ Intro (Conditional Proof)

If a proof contains a subproof with a single hypothesis p and last line q, you may close the subproof and write, on the very next line, the conditional $p \supset q$, with the justification "$\supset$ Intro" (citing the lines of the subproof). Example:

1	$\neg P \wedge (P \wedge R)$	Hyp
2	$P \wedge R$	$\wedge$ Elim 1
3	P	$\wedge$ Elim 2
4	$(\neg P \wedge (P \wedge R)) \supset P$	$\supset$ Intro 1–3

(1.7)

Note that the vertical line indicating the subproof ends just before the line containing the conditional conclusion (4). The hypothesis has been "discharged,"

and the conditional is no longer being asserted merely "under the hypothesis" stated in line 1.

Be careful to add parentheses around the antecedent and consequent of the conditional when needed to avoid ambiguity.

⊃ Elim (Modus Ponens)

If the formulas p and $p \supset q$ both occur in a subproof, you may write down q in the same subproof with the justification "⊃ Elim." Example:

1	P	Hyp
2	$P \supset (S \wedge \neg Q)$	Hyp
3	$S \wedge \neg Q$	⊃ Elim 1, 2

(1.8)

≡ Intro

You prove a biconditional by combining two Conditional Proof subproofs, one in each direction. Example:

1	$P \wedge P$	Hyp
2	P	∧ Elim 1
3	P	Hyp
4	$P \wedge P$	∧ Intro 3, 3
5	$(P \wedge P) \equiv P$	≡ Intro 1–4

(1.9)

≡ Elim

If the formulas p and either $p \equiv q$ or $q \equiv p$ both occur in a subproof, you may write down q in the same subproof with justification "≡ Elim." Example:

1	P	Hyp
2	$P \equiv (S \wedge \neg Q)$	Hyp
3	$S \wedge \neg Q$	≡ Elim 1, 2

(1.10)

⊥ Elim

If ⊥ occurs in a subproof, you may write down any formula in the same subproof, with justification "⊥ Elim." Example:

1	⊥	Hyp
2	$\neg(P \vee Q) \supset R$	⊥ Elim 1

(1.11)

The basic idea: "the absurd" proves anything. (Why, you ask? That is a question we'll return to later.)

¬ Intro

If a proof contains a subproof with hypothesis p and last line ⊥, you may close off the subproof and write, as the very next line, $\neg p$, with the justification "¬ Intro" (citing the lines of the subproof). Example:

1	$\neg P \supset \bot$	Hyp
2	$\neg P$	Hyp
3	⊥	⊃ Elim + Reit, 1, 2
4	$\neg\neg P$	¬ Intro 2–3

(1.12)

Note: We can't get P directly by ¬ Intro, because it is not the negation of the hypothesis (though it is *equivalent* to the negation of the hypothesis). To get P we would need to use the ¬¬ Elim rule (described below).

¬ Elim

If a formula p and its negation $\neg p$ both occur in a subproof, you may write down ⊥ in the same subproof with justification "¬ Elim." Example:

1	$\neg(P \wedge R)$	Hyp
2	$P \wedge R$	Hyp
3	⊥	¬ Elim 1–2

(1.13)

The basic idea is that "the absurd" can be derived directly from any pair of explicitly contradictory formulas.

¬¬ Elim (Double Negation Elimination, DNE)

If a proof contains a $\neg\neg p$, you may write down p, with the justification "¬¬ Elim" (citing the line of the doubly negated formula). Example:

1	$\neg\neg P$	Hyp	(1.14)
2	P	¬¬ Elim 1	

This rule (also called Double Negation Elimination or DNE) is an anomaly in that, unlike the other elimination rules, it removes *two* connectives. We will discuss it further in §6.3.

∨ Intro

If a formula p occurs in a subproof, then you may write down either $p \vee q$ or $q \vee p$ in the same subproof, with the justification "∨ Intro." Example:

1	$\neg Q$	Hyp	(1.15)
2	$P \vee \neg Q$	∨ Intro 1	

∨ Elim (Dilemma)

If a subproof contains a disjunction $p \vee q$ and immediately contains two subproofs, the first hypothesizing p and ending with r, the second hypothesizing q and ending with r, the you may write down r in the same subproof, with justification "∨ Elim."

Here is the pattern and a concrete example:

$p \vee q$
| p
| ⋮
| r
| q
| ⋮
| r
r

1	$A \vee (\neg A \wedge C)$	Hyp	(1.16)
2	$\quad A$	Hyp	
3	$\quad A \vee C$	∨ Intro 2	
4	$\quad \neg A \wedge C$	Hyp	
5	$\quad C$	∧ Elim 4	
6	$\quad A \vee C$	∨ Intro 5	
7	$A \vee C$	∨ Elim 1–6	

1.1.4 Proof strategy

When trying to construct a proof, it is often helpful to start by looking at the main connective of the conclusion, and asking what it would take to obtain it using the introduction rule for that connective. By repeating this process one can often derive the whole proof structure.

For example, suppose we are asked to prove $R \supset ((P \wedge R) \vee (Q \wedge R))$ from $P \vee Q$. The main connective of the conclusion is $\supset$, so we can start by sketching out an application of $\supset$ Intro, leaving lots of space to fill it in:

$$\begin{array}{|l|l|}
\hline
P \vee Q & \text{Hyp} \\
\hline
\quad \begin{array}{|l|l|}
\hline
\quad R & \text{Hyp} \\
\hline
\quad \\
\quad ??? \\
\quad \\
\quad (P \wedge R) \vee (Q \wedge R) & \\
\hline
\end{array} & \\
R \supset ((P \wedge R) \vee (Q \wedge R)) & \supset \text{Intro} \\
\hline
\end{array} \tag{1.17}$$

Our problem is now reduced to that of deriving $(P \wedge R) \vee (Q \wedge R)$ from $P \vee Q$ and R. Can we use $\vee$ Intro? For that we'd need to have one of the disjuncts, either $P \wedge R$ or $Q \wedge R$. There doesn't seem to be any way to get either of these, so we switch gears and try to work forwards, from the premises. We have a disjunction, $P \vee Q$, so the natural thing to try is $\vee$ Elim: if we can derive $(P \wedge R) \vee (Q \wedge R)$ from P, and then from Q, we can derive it from $P \vee Q$.

We can sketch in what this would look like mechanically, leaving space for the "guts" of the proof:

$$
\begin{array}{|l|l}
\hline
P \vee Q & \text{Hyp} \\
\hline
\quad \begin{array}{|l|l}
\hline
\quad R & \text{Hyp} \\
\hline
\quad P \vee Q & \text{Reit 1} \\
\quad \begin{array}{|l|l}
\hline
\quad P & \text{Hyp} \\
\hline
\quad ??? & \\
\quad (P \wedge R) \vee (Q \wedge R) &
\end{array} & \\
\quad \begin{array}{|l|l}
\hline
\quad Q & \text{Hyp} \\
\hline
\quad ??? & \\
\quad (P \wedge R) \vee (Q \wedge R) &
\end{array} & \\
\quad (P \wedge R) \vee (Q \wedge R) & \vee \text{ Elim} \\
R \supset ((P \wedge R) \vee (Q \wedge R)) & \supset \text{ Intro}
\end{array}
$$

(1.18)

Now it just remains to fill in the gaps marked ???. $(P \wedge R) \vee (Q \wedge R)$ is a disjunction, so the first thing to try is $\vee$ Intro. Can we get either of its disjuncts from P? Yes:

$$
\begin{array}{|l|l|l}
\hline
\quad P & \text{Hyp} & \\
\hline
\quad R & \text{Reit 2} & \\
\quad P \wedge R & \wedge \text{ Intro} & \\
\quad (P \wedge R) \vee (Q \wedge R) & \vee \text{ Intro} &
\end{array}
$$

(1.19)

We leave it to the reader to complete the other gap in proof (1.18) and fill in the line numbers.

1.1.5 The relation of semantics and proofs

Once we have a semantics and a proof system for our logic, we can ask questions about how they are related. Ideally, we'd like to have the following two properties:

> Our system is *sound* if, whenever q can be proved from hypotheses $p_1, \ldots, p_n$ in our proof system, q is a logical consequence of $p_1, \ldots, p_n$.
>
> Our system is *complete* if, whenever q is a logical consequence of $p_1, \ldots, p_n$, q can be proved from hypotheses $p_1, \ldots, p_n$ in our system.

Exercise 1.2: Deductions and invalidity

1. For each of the following arguments, either show that it is valid by giving a proof in our system, or show that it is invalid by describing a model on which the premises are true and the conclusion false:

$$\text{a)}\ \frac{A \qquad B \supset (A \supset B)}{B} \qquad \text{b)}\ \frac{A \equiv (B \vee C) \qquad A \vee B}{A} \qquad \text{c)}\ \frac{A \vee (B \supset C) \qquad B}{A \vee C}$$

2. I've just completed a correct deduction of $\neg p$ from q in a sound deduction system. Can I conclude that q does not truth-functionally imply p? Why or why not?

In fact, our proof system does have both these properties relative to our semantics. But this is not just obvious. It is something that has to be proved. (If you take a course in metalogic, you can find out how this is done.)

1.2 Predicate logic

The inference

$$(5)\quad \frac{\text{Felix is a cat}}{\text{Something is a cat}}$$

is intuitively valid. But from the point of view of propositional logic, we can only represent it as

$$(6)\quad \frac{F}{S}$$

which is invalid. To capture the validity of (5), we need to be able to represent the way in which sentences are composed out of names, predicates, and quantifiers, as well as the sentential connectives. That calls for some new syntax.

1.2.1 Grammar

- A *variable* is a lowercase w, x, y, or z, possibly with a numerical subscript (x_1, y_{14}, etc.).
- An *individual constant* is a lowercase letter other than w, x, y, or z, possibly with a numerical subscript.
- A *term* is an individual constant or a variable.
- A *predicate* is a capital letter other than W, X, Y, or Z, possibly with a numerical subscript. Predicates can be classified as one-place, two-place, and in general n-place, depending on how many argument places they have.
- A *formula* is any of the following:
  - $\bot$
  - An *atomic formula*—an n-place predicate followed by n terms. (For example: Fxy, G_1a_{15}.)
  - $\forall\alpha\phi$ or $\exists\alpha\phi$, where α is a variable and ϕ is a formula.
  - $\neg\phi$, where ϕ is a formula.
  - $(\phi \vee \psi)$, $(\phi \wedge \psi)$, $(\phi \supset \psi)$, or $(\phi \equiv \psi)$, where ϕ and ψ are formulas.

  Nothing else is a formula.

Here we use Greek letters ϕ, ψ, χ, and ξ as *metavariables* ranging over formulas. Similarly, α and β range over terms, and Φ and Ψ range over predicates.[4] They are called *meta*variables because, like 'p' and 'q' in §1.1, they are not part of the logical language we are describing (the *object language*), but the language we use to talk about it (the *use language* or *metalanguage*).

The operators $\forall\alpha$ and $\exists\alpha$ are called *quantifiers*.[5] $\forall\alpha$ is the *universal* quantifier and may be read "everything is such that ...." $\exists\alpha$ is the *existential* quantifier and may be read "at least one thing is such that ...." In translating from predicate logic to English, we may start with these renderings and then find more colloquial ones. For example:

[4] See Appendix A for a guide to pronouncing the Greek letters used here.

[5] You may have seen different notation: sometimes (x) is used instead of $\forall x$, and sometimes the quantifier is surrounded by parentheses, as in $(\forall x)$.

(7) $\forall x(Fx \supset Gx)$
Everything is such that if it is F, then it is G.
All Fs are Gs.[6]

(8) $\neg\exists x(\forall yRxy)$
It is not the case that at least one thing (x) is such that everything (y) is such that it (x) Rs it (y).
It is not the case that at least one thing is such that it Rs everything.
There isn't anything that Rs everything.
Nothing Rs everything.

Note that $\exists x\exists xFx$ is a formula in our system, as is $\exists xFa$. Such formulas may have been disallowed in the logical system you learned, but our semantics for quantifiers (§1.2.3) gives them a clear interpretation.

1.2.2 Scope

The *scope* of a quantifier is the formula directly following the quantifier:

- In $\forall x(Fx \supset Gx)$, the scope of the quantifier is the formula $(Fx \supset Gx)$.
- In $\forall xFx \supset Gx$, the scope of the quantifier is the formula Fx. (Remember, $Fx \supset Gx$ is not a formula, because it lacks the required parentheses. We may omit these *only* at the outer level, as a convenience.)
- In $\forall x\neg Fx \vee Ga$, the scope of the quantifier is the formula $\neg Fx$.

A quantifier $\exists\alpha$ or $\forall\alpha$ will *bind* all occurrences of α within its scope, except those that are already bound by other quantifiers. A variable that is not *bound* by a quantifier is called *free*. A formula containing free variables is called an *open formula*. A formula without free variables is called a *closed formula* or *sentence*.

1.2.3 Semantics

In §1.1, we described a model as something that provides enough information to determine truth values for all of the formulas in a language. (In propositional logic, this is just an assignment of truth values to the propositional constants.) Now we must qualify that slightly: a model must determine truth values for all of the *closed* formulas (sentences) in a language. Open formulas do not have truth values.

A *model* for our language of predicate logic consists in

[6] Although this is a standard rendering, you might find yourself wondering whether these last two lines are really equivalent. We will revisit this issue in Chapters 2 and 4.

Exercise 1.3: Translations

1. Translate the following into logical notation. Provide a "dictionary" that associates individual constants and predicate letters with English names and predicates, and be sure to specify a domain.
   a) There's a woman who adopts every cat she meets.
   b) Not all cats and dogs are furry.
   c) Every dog despises at least one cat that has scratched one of its (the dog's) friends.
2. Translate the following into English (provide a dictionary—you may make it up):
   a) $\neg\exists x(Lx \wedge \forall y(Py \supset Sxy))$
   b) $\forall x((Fx \wedge \forall y(Gy \supset Hxy)) \supset \exists z(Cz \wedge Lxz))$
3. In each of the following sentences, circle the free variables and draw arrows from each of the bound variables to the quantifier that binds it.
   a) $\forall x(Fy \supset Gxy) \supset Gyx$
   b) $\forall x\exists y(Gxy \supset \exists xGyx)$
   c) $\forall x(Fy \wedge \exists xFx)$

- a nonempty set of objects—the *domain*, and
- an *interpretation function*, which assigns an interpretation to each individual constant and predicate letter. More specifically, it maps
  - each individual constant to an object in the domain
  - each one-place predicate letter to a set of objects in the domain
  - each two-place predicate letter to a set of ordered pairs of objects in the domain
  - each n-place predicate letter ($n > 2$) to a set of ordered n-tuples of objects in the domain.

	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$
D	{1, 2, 3, Paris}	the set of integers	$\{x : x$ has played basketball$\}$
$I(F)$	$\{1, 3\}$	$\{x : x > 0\}$	$\{x : x$ is Chinese$\}$
$I(G)$	$\{\langle 1, 2\rangle, \langle 3, 3\rangle\}$	$\{\langle x, y\rangle : x > y\}$	$\{\langle x, y\rangle : x$ is taller than $y\}$
$I(a)$	Paris	1	Michael Jordan

Table 1.1: Three models specified using set-theoretic notation. Here F is a one-place predicate, G is a two-place predicate, and a is an individual constant. The formula Fa is true in $\mathscr{M}_2$ (since $1 > 0$) but not in $\mathscr{M}_1$ (Paris is not a member of the set $\{1, 3\}$) or $\mathscr{M}_3$ (Michael Jordan is not Chinese). The formula $\forall x(Fx \supset \exists y Gxy)$ is true in $\mathscr{M}_2$, since every integer greater than 0 is greater than some integer. Is it true in $\mathscr{M}_1$? What would you need to know in order to know if it is true in $\mathscr{M}_3$?

In specifying a model, we'll generally only write down the interpretations of individual constants and predicate letters that are relevant for our purposes, in the same way that we omitted irrelevant propositional constants when giving models for propositional logic.

Table 1.1 gives some examples of models, using set-theoretic notation. (If you are not familiar with this notation, see Appendix B for a quick primer.) We can also specify models informally using pictures, as illustrated in Fig. 1.1.

We say that a sentence (closed formula) is *true in a model* just in case it is true when the quantifiers are interpreted as ranging over objects in the domain (and no others) and the individual constants and predicate letters are interpreted as having just the extensions assigned to them by the interpretation function. (The *extension* of an individual constant is the object it refers to; the extension of a predicate is the set of objects it is true of, or for a relation, the set of tuples of objects that satisfy it.)

To state this condition precisely, we need to define truth in a model for each type of formula, as we did for propositional logic in §1.1.2. But here we hit a snag. Truth in a model is defined only for closed formulas: an open formula such as Fxa cannot be said to be true or false in a model, because a model only interprets predicates and individual constants, not variables. But quantified formulas like $\forall x Fxa$ have open formulas as their parts. So we cannot do what we did before, defining truth in $\mathscr{M}$ for each formula in terms of truth in $\mathscr{M}$ for its constituents.

The solution to this problem (due to Tarski 1935) is to start by defining truth in a model on an assignment of values to the variables, and then define truth in a model in terms of this. An *assignment* is a function that maps each variable to an object in the domain. To avoid verbiage, we will adopt the following abbreviations:

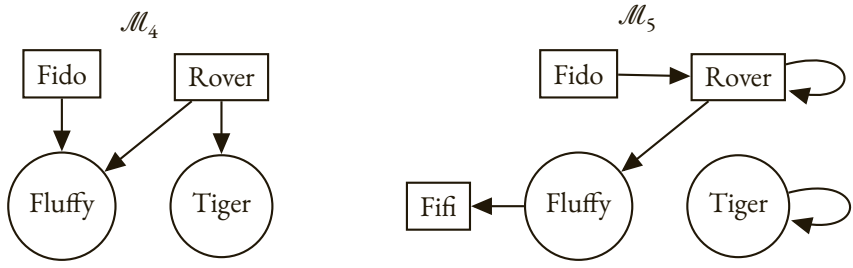


Figure 1.1: Two models given pictorially. Here the rectangles are the Ds and the arrows represent the C relation. The formula $\forall x(Dx \supset \exists y Cxy)$ is true in $\mathcal{M}_4$ but not in $\mathcal{M}_5$. To see this, it may help to interpret Dx as 'x is a dog', and Hxy as 'x chases y'. Try evaluating the formulas $\exists x \exists y(Dx \wedge \neg Dy \wedge Cxy)$ and $\exists x Cxx$ in both models. Unless noted otherwise, it is assumed in such diagrams that the domain comprises just the objects pictured.

$\vDash^{v}_{\mathcal{M}} \phi$	ϕ is true in the model $\mathcal{M}$ on the assignment v
$\nvDash^{v}_{\mathcal{M}} \phi$	ϕ is not true in the model $\mathcal{M}$ on the assignment v
$[\![\alpha]\!]^{v}_{\mathcal{M}}$	$= v(\alpha)$ if α is a variable
	$= I(\alpha)$ if α is an individual constant.

We can now specify what it is for an arbitrary formula ϕ (open or closed) to be true in a model $\mathcal{M} = \langle D, I \rangle$ on an assignment v.

- If ϕ is an atomic formula $\Phi\alpha_1 \ldots \alpha_n$, where Φ is an n-place predicate and $\alpha_1 \ldots \alpha_n$ are terms, $\vDash^{v}_{\mathcal{M}} \phi$ iff $\langle [\![\alpha_1]\!]^{v}_{\mathcal{M}}, \ldots, [\![\alpha_n]\!]^{v}_{\mathcal{M}} \rangle \in I(\Phi)$.
- If ϕ is $\bot$, then $\nvDash^{v}_{\mathcal{M}} \phi$.
- If ϕ is $\neg\psi$, then $\vDash^{v}_{\mathcal{M}} \phi$ iff $\nvDash^{v}_{\mathcal{M}} \psi$.
- If ϕ is $(\psi \wedge \chi)$, then $\vDash^{v}_{\mathcal{M}} \phi$ iff $\vDash^{v}_{\mathcal{M}} \psi$ and $\vDash^{v}_{\mathcal{M}} \chi$.
- If ϕ is $(\psi \vee \chi)$, then $\vDash^{v}_{\mathcal{M}} \phi$ iff $\vDash^{v}_{\mathcal{M}} \psi$ or $\vDash^{v}_{\mathcal{M}} \chi$.
- If ϕ is $(\psi \supset \chi)$, then $\vDash^{v}_{\mathcal{M}} \phi$ iff $\nvDash^{v}_{\mathcal{M}} \psi$ or $\vDash^{v}_{\mathcal{M}} \chi$.
- If ϕ is $(\psi \equiv \chi)$, then $\vDash^{v}_{\mathcal{M}} \phi$ iff either $\vDash^{v}_{\mathcal{M}} \psi$ and $\vDash^{v}_{\mathcal{M}} \chi$ or $\nvDash^{v}_{\mathcal{M}} \psi$ and $\nvDash^{v}_{\mathcal{M}} \chi$.
- If ϕ is $\forall\alpha\psi$, where α is a variable, then $\vDash^{v}_{\mathcal{M}} \phi$ iff for *every* assignment v' that agrees with v on the values of every variable except possibly α, $\vDash^{v'}_{\mathcal{M}} \psi$.

(The idea here is that we want to consider every way of assigning a value to α. This means looking at multiple assignments. But we only want to shift the value of α, not other variables, so we only look at assignments that agree with v on all variables other than α.)

- If ϕ is $\exists\alpha\psi$, where α is a variable, then $\vDash^{v}_{\mathcal{M}} \phi$ iff for *some* assignment v' that agrees with v on the values of every variable except possibly α, $\vDash^{v'}_{\mathcal{M}} \psi$.

Having defined the condition for any open or closed formula ϕ to be true in a model $\mathcal{M}$ on an assignment v, we can define truth in a model (not relativized to an assignment) for *closed* formulas as follows:

A closed formula ϕ is *true in a model* $\mathcal{M}$ iff for every assignment v, $\vDash^{v}_{\mathcal{M}} \phi$.[7]

Once we have defined truth in a model in this way, we can define logical consequence, logical truth, logical equivalence, logical independence, and so on by quantifying over models, just as we did for propositional logic (see §1.1.2). The only difference is that the models we quantify over are now more complicated. Thus, for example, to show that an argument is invalid, we now need to find a domain D and interpretation I such that the premises are true in D, I but the conclusion is false in D, I.

1.2.4 Proofs

All of the rules for propositional logic can be used in predicate logic as well, but we need a few new rules to deal with quantifiers.

Substitution instances

A *substitution instance* of a quantified formula is the result of deleting the quantifier and its associated variable, then replacing every variable bound by the quantifier with the same individual constant. Thus, for example, $Faab$ is a substitution instance of $\exists xFxxb$ (replace every x with a) and also of $\forall yFyab$ (replace every y with a), but *not* of $\forall xFxax$.

∀ Elim (Universal Instantiation)

You may write down any substitution instance of any universally quantified formula that occurs in the same subproof, with the justification "∀ Elim" (citing the line containing the quantified formula). Example:

[7] We could have said "some assignment" instead of "every assignment"; it doesn't matter, because if ϕ is a closed sentence, its truth won't vary from one assignment to the next.

Exercise 1.4: Semantics for predicate logic

1. For each of the three sample models in Table 1.1, above, say which of the following sentences are true in that model:
   a) $\exists x(Fx \wedge Gxa)$
   b) $\exists x \exists y(Gxy \wedge Gyx)$
   c) $\exists x \forall y \neg Gyx$
2. Complete the definitions, using the first line as a paradigm:
   a) A sentence is *logically true* iff it is true in all models.
   b) A sentence is *logically false* iff ...
   c) Two sentences are *logically equivalent* iff ...
   d) One sentence *logically implies* another iff ...
   e) A sentence (S) is a *logical consequence* of a set of sentences (Γ) iff ...
   f) An argument is *logically valid* iff ...
3. Use models to show the following:
   a) $\exists x \forall y Fxy$ and $\forall y \exists x Fxy$ are not logically equivalent.
   b) $(Fa \supset \forall x Fx) \supset Fb$ is not a logical truth.
   c) $Fa \wedge Gb$ does not logically imply $\exists x(Fx \wedge Gx)$.

1	$\forall x \exists y Fxay$	Hyp	
2	$\exists y Faay$	$\forall$ Elim 1 a/x	(1.20)
3	$\exists y Fbay$	$\forall$ Elim 1 b/x	

Notes:

1. It is a very good habit to indicate which constant is replacing which variable, as in the example.

2. There are no restrictions on which individual constant you use. Just be sure you replace *every* occurrence of the bound variable with the same constant. You can't use ∀ Elim to go from $\forall x Fxax$ to $Fbax$, because not every occurrence of the bound variable x was replaced by b.

3. A universally quantified formula is a formula whose *main connective* is ∀. You can't use ∀ Elim to go from $\forall x Fx \vee \forall x Gx$ to $Fa \vee Ga$, because the former is not a universally quantified formula (the main connective is ∨).

∃ Intro (Existential Generalization)

If a substitution instance of an existentially quantified formula occurs in a subproof, you may write down the existentially quantified formula in the same subproof, with the justification "∃ Intro" (citing the line containing the instance). Example:

1	$\forall y Faya$	Hyp
2	$\exists x \forall y Fxya$	∃ Intro 1 a/x
3	$\exists x \forall y Fxyx$	∃ Intro 1 a/x

(1.21)

Notes:

1. An existentially quantified formula is a formula whose *main connective* is ∃. $\exists x Fx \vee Ga$ is not an existentially quantified formula, and it can't be obtained by ∃ Intro from $Fb \vee Ga$.

2. Whereas with ∀ Elim you move from a quantified formula to an instance, with ∃ Intro you move from an instance to a quantified formula.

3. Line 1 above is an instance of both 2 and 3.

∀ Intro (Universal Generalization)

You can derive a universally quantified formula $\forall \alpha \phi$ from a subproof whose last step is a substitution instance, with an individual constant in place of α, and whose first step is a *flagging step* containing that individual constant in a box. The justification is "∀ Intro" (citing the lines of the subproof).

A *flagging step* is like a hypothesis, but instead of a formula, it consists of an individual constant in a box:

1	$\boxed{a}$

There is one important restriction:

Flagging restriction The flagged constant may not occur outside of the subproof where it is introduced.

So pick a constant that does not occur in the premises or conclusion or in any previous flagging step.

The flagging step is a formal representation of "Take an arbitrary individual—call it Joe." We then argue that Joe has such and such a property, and since Joe was arbitrary, the same could be shown about any object. The flagging restrictions are there to make sure the individual is really arbitrary, not one that you have already said something about elsewhere in the proof.[8]

Example:

1	$\forall x(Gx \supset Hx)$	Hyp
2	$\forall x(Hx \supset Fx)$	Hyp
3	$\boxed{a}$	
4	$Ga \supset Ha$	$\forall$ Elim + Reit 1 a/x
5	$Ha \supset Fa$	$\forall$ Elim + Reit 2 a/x
6	Ga	Hyp
7	Ha	$\supset$ Elim 4 + Reit 6
8	Fa	$\supset$ Elim 5 + Reit 7
9	$Ga \supset Fa$	$\supset$ Intro 6–8
10	$\forall x(Gx \supset Fx)$	$\forall$ Intro 3–9 a/x

(1.22)

∃ Elim (Existential Instantiation)

If an existentially quantified formula occurs in a subproof, you may start a new subproof with an instance as a hypothesis and the instantial constant "flagged" in a box. You can close the new subproof at any point where you have a formula not containing the flagged constant. This final formula may then be written outside the subproof, with justification "∃ Elim", citing the existentially quantified formula and the subproof. As before, the flagged constant may not occur outside

[8] You may have learned a deduction system that does not require a subproof for ∀ Intro, instead allowing $\forall F(x)$ to be inferred directly from any instance $F(c)$, provided the constant c is not used in any undischarged assumptions. But requiring a subproof with a flagged constant makes the constant's role as denoting an arbitrary individual more explicit.

of the subproof where it is introduced. Example:

1	$\exists x(Gx \wedge Ha)$		
2	$\quad Gb \wedge Ha$	$\boxed{b}$	b/x
3	$\quad Gb$		$\wedge$ Elim 2
4	$\quad \exists xGx$		$\exists$ Intro 3
5	$\exists xGx$		$\exists$ Elim 1, 2–4

(1.23)

Notes:

1. We could not have closed off the subproof after line 3, since the flagged constant cannot occur in the last line of the main proof.
2. We could not have used *a* as our flagged term in line 2, since it occurs in line 1.

Substitution rules

The introduction and elimination rules for the quantifiers and propositional connectives, together with the structural rules, give us all we need for a complete proof system. But to make quantificational proofs less tedious, we will also allow the use of two more rules. Unlike the rules we have seen so far, these are *substitution rules*, which allow one formula to be substituted for another, even if it just part of a larger formula.

QNE (Quantifier-Negation Equivalences)

You may use the following substitution rules at any point in a proof, citing "QNE" and the line number as justification. They are all reversible. (See the examples to follow.)

$$\neg\forall x\phi \iff \exists x\neg\phi$$
$$\neg\exists x\phi \iff \forall x\neg\phi$$

Examples:

1	$\neg\exists x(Gx \wedge Ha)$	
2	$\forall x\neg(Gx \wedge Ha)$	QNE 1

(1.24)

1	$Ha \supset \forall x\neg Gx$	
2	$Ha \supset \neg\exists xGx$	QNE 1

(1.25)

Note that QNE is applied to a subformula in example (1.25). The main connective in (1) is '⊃,' not a quantifier. That's okay, because the QNE rules are substitution rules, not rules of inference.

Taut Equiv (Tautological Equivalence)

What if you wanted to derive $\forall x(Gx \supset \neg Hx)$ from $\neg\exists x(Gx \wedge Hx)$? Given the rules we have so far, you'd have to take a circuitous path:

1	$\neg\exists x(Gx \wedge Hx)$	
2	$\forall x\neg(Gx \wedge Hx)$	QNE 1
3	$\boxed{b}$	
4	$\neg(Gb \wedge Hb)$	∀ Elim + Reit 2, b/x
5	Gb	Hyp
6	Hb	Hyp
7	$Gb \wedge Hb$	∧ Intro + Reit 5, 6
8	$\perp$	¬ Elim + Reit 4, 7
9	$\neg Hb$	¬ Intro 6–8
10	$Gb \supset \neg Hb$	⊃ Intro 5–9
11	$\forall x(Gx \supset \neg Hx)$	∀ Intro 3–10, b/x

(1.26)

To simplify this kind of proof, we introduce a new substitution rule, Taut Equiv, that allows you to *substitute* truth-functionally equivalent formulas for each other, even when they occur embedded inside quantifiers or other operators. Then we can do:

1	$\neg\exists x(Gx \wedge Hx)$	
2	$\forall x\neg(Gx \wedge Hx)$	QNE 1
3	$\forall x(Gx \supset \neg Hx)$	Taut Equiv 2

(1.27)

We'll allow Taut Equiv only in proofs involving quantifiers.

1.3 Identity

The following inference seems valid:

Exercise 1.5: Deductions for predicate logic

1. Use Fitch-style natural deductions to prove the following theorems:
   a) $\forall x((Fx \wedge Gx) \supset Fx)$
   b) $\neg\exists x(Fx \wedge Gx) \supset (\forall xFx \supset \neg\exists xGx)$
   c) $\exists x\forall y\forall zFxyz \supset \forall y\forall z\exists xFxyz$
2. Use Fitch-style natural deductions to prove

$$\begin{array}{ll} & \exists x(Px \wedge Sx) \\ \text{a)} & \forall x(Sx \supset Rxb) \\ \hline & \exists xRxb \end{array} \qquad \begin{array}{ll} & \forall x(Px \supset \exists yFyx) \\ \text{b)} & \forall x\forall y(Fyx \supset Lyx) \\ \hline & \forall x(Px \supset \exists yLyx) \end{array}$$

3. Use Fitch-style natural deductions to prove $\exists x\neg Px$ from $\neg\forall xPx$ without using the QNE rules.

$$\begin{array}{ll} & \text{There is at most one cat that is black} \\ (9) & \text{There are at least two cats} \\ \hline & \text{There is at least one cat that is not black} \end{array}$$

However, we cannot capture its validity using just the resources of basic predicate logic. To represent its premises and conclusion, we will need to introduce a sign for *identity*.

In ordinary language, when we say that two shirts are identical, we mean that they are the same color, style, fit, and so on. In logic, the term 'identity' is used for *numerical identity*: to say that A is identical to B is to say that they are the same object. In the logical sense, Clark Kent and Superman are identical, but Clark Kent is not identical with his twin who looks just like him.

As we will see, adding a sign for identity to predicate logic increases its expressive power, allowing us to say things we couldn't have said without it. Without an identity sign, for example, we can't say that there are at least two things that are F. $\exists x\exists y(Fx \wedge Fy)$ can be true even if there's just one object in the domain that is F. To say that there are at least two, we need to be able to say that x and y are not the same: $\exists x\exists y(Fx \wedge Fy \wedge \neg x{=}y)$.

1.3.1 Grammar

The identity sign (=) is a two-place predicate. By convention, we write one argument on the left and one on the right (as in $a{=}b$). We should not let this convention obscure the fact that grammatically = is just a two-place predicate, like the G in Gab. We could just as well have written $=ab$ or Iab.

Sometimes the nonidentity sign (≠) is also used. We can introduce it as a defined term:

Nonidentity $\alpha \neq \beta \equiv \neg(\alpha{=}\beta)$

1.3.2 Semantics

The extension of = in a model is the relation each thing bears to itself and to no other thing (the identity relation). For example, if the domain is {1, 2, 3}, then the extension of = is $\{\langle 1,1\rangle, \langle 2,2\rangle, \langle 3,3\rangle\}$. That is not to say that all true identity statements are the tautologous kind ($a{=}a$). $a{=}b$ can be true in a model, provided that a and b get assigned the same interpretation (the same object) in that model.

1.3.3 Proofs

To do proofs with identity, we'll need two new rules.

= Intro (Reflexivity) Where α is an individual constant, you may write $\alpha{=}\alpha$ on any line of a proof, with justification "= Intro."

= Elim (Substitution of Identicals) From premises $\alpha{=}\beta$ (or $\beta{=}\alpha$) and ϕ, where α and β are individual constants and ϕ a sentence containing α, you may conclude any formula ψ that results from ϕ by replacing one or more occurrences of α with β.

Note that *both* = Elim steps in the following proof are valid:

1	$\exists x Raxa$	Hyp
2	$a{=}b$	Hyp
3	$\exists x Rbxb$	= Elim 1, 2
4	$\exists x Rbxa$	= Elim 1, 2

(1.28)

(3) is a valid step because it is the result of substituting b for both occurrences of a in line (1). (4) is a valid step because it is the result of substituting b for the first occurrence of a in (1).

This is all you need for proofs with identity. Here's an example.

1	$\exists x(Fx \wedge Gxb)$	Hyp	
2	$a{=}b$	Hyp	
3	$\exists x(Fx \wedge Gxa)$	= Elim 1, 2	
4	$Fc \wedge Gca$ $\boxed{c}$	3 c/x	
5	Gca	$\wedge$ Elim 4	(1.29)
6	$c{=}c$	= Intro	
7	$Gca \wedge c{=}c$	$\wedge$ Intro 5,6	
8	$\exists x(Gxa \wedge x{=}x)$	$\exists$ Intro 7 c/x	
9	$\exists x(Gxa \wedge x{=}x)$	$\exists$ Elim 3, 4–8	

1.4 Use and mention

The word 'sentence' is used to say things about sentences. But when I say

(10) The word 'sentence' has eight letters.

I am not saying anything about sentences; I'm talking instead about the *word* 'sentence', which I am not *using* but *mentioning*.

Here I have used the convention (introduced by Frege 1893) of putting a phrase in single quotation marks to indicate that one is mentioning it. This is just one of several conventions one might adopt. Some authors use italics to indicate mention. And others just use the words *autonymously*—that is, as names of themselves (Church 1956, §08; Carnap 2002, §42)—and leave it to the reader to figure out from context when they are functioning as names of themselves and when they are being used in their normal way. That is what we have done so far in this chapter, and it works well when the language being used is different from the language being discussed.

However, we will soon be discussing some issues where use/mention ambiguities can lead to fallacious reasoning. So we will start being more explicit, using single quotation marks to indicate mention. Thus, instead of writing, confusingly,

(11) a. Boston is a city. Boston is the name of a city.
b. An hour is longer than a minute, but minute is longer than hour.
c. An expressively complete logic can contain either and and not or or and not.

Exercise 1.6: Identity

1. How would you express the following in predicate logic with identity?
   a) Every logician loves someone other than herself.
   b) The only one who respects Richard is Sue.
   c) There are at least two rich dogs.
   d) There are at most two smart dogs.
   e) Liz is the tallest spy.
   f) Liz is the tallest rider who roped at least two calves.
2. (a) Give a formula, using quantifiers and identity, that is true in every model with a domain of one object and false in some model with a domain of two objects. (b) Give a formula, *not* using quantifiers or identity, that has this property.
3. Without an identity sign you can't produce a sentence that says that there are at least two Fs. However, you *can* produce a sentence without identity that is only true in models whose domains contain at least two things that fall into the extension of F. Can you find one?
4. Prove that the following rules are valid. (Give a deduction.) Once you have done this, you may use these derived rules to simplify proofs with identity.

   Symmetry $\dfrac{a=b}{b=a}$ *Transitivity* $\dfrac{a=b \quad b=c}{a=c}$

5. Prove $Fa \equiv \exists x(Fx \wedge x=a)$.
6. Suppose that you have a quantifier $\exists^n x$, meaning "there are at least n x..." How could you define $\exists^{n+1} x$ in terms of $\exists^n x$?
7. Translate argument (9) from the beginning of this section, and give a deduction to show that it is valid.
8. ⋆The identity sign is treated differently from other predicates in first-order logic. Can you think of any reasons for this?

we will write

(12) a. Boston is a city. 'Boston' is the name of a city.
b. An hour is longer than a minute, but 'minute' is longer than 'hour'.
c. An expressively complete logic can contain either 'and' and 'not' or 'or' and 'not'.

In giving semantic clauses for logical connectives and operators, we have used phrasing like this:

(13) Where ϕ and ψ are formulas, $\phi \wedge \psi$ is true in a model $\mathscr{M}$ iff ϕ is true in $\mathscr{M}$ and ψ is true in $\mathscr{M}$.

How can we rephrase this in a way that is more careful about use and mention? Well, we might try:

(14) Where ϕ and ψ are formulas, '$\phi \wedge \psi$' is true in a model $\mathscr{M}$ iff 'ϕ' is true in $\mathscr{M}$ and 'ψ' is true in $\mathscr{M}$.

But this won't work! Remember, ϕ and ψ are variables whose *values* are formulas. They are not formulas themselves. The expression

(15) '$\phi \wedge \psi$'

denotes the sequence of symbols:

(16) $\phi \wedge \psi$

which is not, itself, a formula of the language we are describing.

How can we say what we want to say, then? Well, we could say this:

(17) Where ϕ and ψ are formulas, the formula consisting of ϕ concatenated with '$\wedge$' concatenated with ψ is true in a model $\mathscr{M}$ iff ϕ is true in $\mathscr{M}$ and ψ is true in $\mathscr{M}$.

We could make this simpler by introducing a notation for concatenation ('$\frown$'). But the result is still pretty ugly:

(18) Where ϕ and ψ are formulas, $\phi \frown$ ' $\wedge$ ' $\frown \psi$ is true in a model $\mathscr{M}$ iff ϕ is true in $\mathscr{M}$ and ψ is true in $\mathscr{M}$.

For this reason, W. V. O. Quine (1940) invented a device of *quasiquotation* or *corner quotes*.[9] Using corner quotes, we can write our semantic clause like this:

[9] Quine's device is used not just in philosophical logic, but in programming languages: Lisp, for example, contains dedicated syntax for quasiquotation.

(19) Where ϕ and ψ are formulas, $\ulcorner \phi \wedge \psi \urcorner$ is true in a model $\mathscr{M}$ iff ϕ is true in $\mathscr{M}$ and ψ is true in $\mathscr{M}$.

You can think of corner quotes as a notational shortcut:

(20) $\ulcorner \phi \wedge \psi \urcorner$

means just the same as

(21) $\phi \frown$ ‘ $\wedge$ ’ $\frown \psi$

More intuitively, you can understand the corner quote notation as follows. A corner-quote expression always denotes another expression, relative to an assignment of values to its variables. To find out what expression it denotes on a given assignment of values to these variables, first replace the variables inside corner quotes with their values (which should be expressions), then convert the corner quotes to regular quotes. So, for example, when ϕ = ‘Cats are furry’ and ψ = ‘Snow is black’, $\ulcorner (\phi \wedge \psi) \urcorner$ = ‘(Cats are furry $\wedge$ Snow is black)’.

Suppose the baby is learning to talk, and says, ‘I like blue,’ ‘I like red,’ ‘I like white,’ ‘I like green,’ and so on for all the color words she knows. To report what happened in a compact way, you might say something like:

(22) For every color word *C*, she said, ‘I like *C*.’

Strictly speaking, though, this reports her as having said, ‘I like *C*’, not ‘I like blue’, etc. To get the desired reading, you can use corner quotes:

(23) For every color word *C*, she said, $\ulcorner$I like *C*$\urcorner$.

which means just the same as:

(24) For every color word *C*, she said, ‘I like’ $\frown$ *C*.

Further readings

For an excellent course in the fundamentals of first-order logic, with exercises, I recommend Nuel Belnap’s unpublished *Notes on the Art of Logic* (Belnap 2009).

Exercise 1.7: Quotation and quasiquotation

1. Add quotation marks where they are needed in the following sentences to mark mention:
   a) Word is a four-letter word.
   b) Boston denotes the name Boston, which denotes the city Boston.
   c) We substitute $a + 3$ for x, if $a + 3$ is a prime number. (Carnap 2002, §42)
2. Rewrite the following using corner-quote notation:
   a) $\phi \frown$ ‘ + ’ $\frown \psi \frown$ ‘=’ $\frown \psi$
   b) ‘∀’ $\frown \alpha \frown \phi$
3. Rewrite the following using regular quotes and the concatenation sign:
   a) $\ulcorner \exists \alpha(\phi \wedge \psi) \urcorner$
   b) $\ulcorner \phi \supset \phi \urcorner$
   c) $\ulcorner \phi \urcorner$
4. Write the expression denoted by the following terms, under an assignment of ‘Fx’ to ϕ, ‘$(Fx \supset Gx)$’ to ψ, and ‘x’ to α:
   a) $\ulcorner \phi \wedge \psi \urcorner$
   b) $\ulcorner \forall \alpha(\psi \supset \phi) \urcorner$

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2 Quantifiers

One might get the impression from a first logic course that the quantifiers ∀ and ∃ suffice for expressing any quantificational generalizations we might like to make. That is not the case. In this chapter, we will explore some ways to extend the quantificational system of basic predicate logic, opening up new expressive possibilities. We will see that quantifiers are better understood as *binary* rather than *unary* operators, and that not all binary quantifiers can be defined in terms of unary ones. We will see how definite descriptions, such as 'the first dog born at sea', can be understood as quantifiers. We will look at *second-order logic*, which allows quantifying into the grammatical positions occupied by predicates. Finally, we will look at the substitutional interpretation of the quantifiers, exploring its motivations and difficulties.

2.1 Beyond ∀ and ∃

2.1.1 What is a quantifier?

We know that '∀' and '∃' are quantifiers, but just what is a quantifier? Webster's *Ninth New Collegiate Dictionary* gives two definitions:

> *a.* a prefixed operator that binds the variables in a logical formula by specifying their quantity. *b.* a limiting noun modifier (as *five* in "the five young men") expressive of quantity and characterized by occurrence before the descriptive adjectives in a noun phrase.

Using definition (b) as a paradigm, we can generate some clear examples:

five young men	*no* young men
all young men	*some* young men
two young men	*more than six* young men
at most four young men	*a few* young men
many young men	*most* young men
a young man	*almost all* young men

Grammatically, these quantifiers are *determiners*—words that determine or limit a noun phrase. (Other determiners in English include 'the', 'my', 'which', and 'those'.) To form a sentence using a determiner, one needs to add *two* things: a noun phrase and a verb phrase:[1]

Det	NP	VP
These	young men	sang in harmony
At least two	women	went swimming
Most	hairless dogs	are cold

In this respect, our natural-language quantifiers are different from the familiar *unary quantifiers* of first-order logic, which just require you to add *one* thing (an open formula) to get a sentence. We can better capture the grammatical form of English sentences using *binary quantifiers*: quantifiers that take *two* open formulas and form a sentence:

$$\begin{aligned} \textit{all}_x(Fx, Gx) &\quad \text{All } F\text{s are } G. \\ \textit{at-least-two}_x(Fx, Gx) &\quad \text{At least two } F\text{s are } G. \\ \textit{most}_x(Fx, Gx) &\quad \text{Most } F\text{s are } G. \end{aligned}$$

You might wonder: why did we not do it this way from the start? In fact, we did. Aristotelian logic used four binary quantifiers:

A	All F are G	I	Some F are G
E	No F are G	O	Some F are not G

Here F and G are *terms* (noun phrases) which can serve as either subjects or predicates. This was the orthodox way of thinking about quantifiers until the end of the nineteenth century, when Gottlob Frege, Bertrand Russell, Giuseppe Peano, and other mathematical logicians introduced unary quantifiers.

The new quantifiers were unquestionably an advance, but it is important to see why. They departed from the Aristotelian quantifiers in two ways:

(i) They take one argument, not two.

(ii) Their arguments are open formulas, not terms.

The key advance was (ii). By allowing quantifiers to apply to arbitrary open formulas, Frege showed us how to express complex mathematical concepts in a way that makes them amenable to logical manipulation. For example, we can

[1] The noun can sometimes be omitted, when it is clear from context: "Most went."

express the idea that an ordering is *dense* (for any two points in the ordering, there is another point between them) as follows:

$$\forall x \forall y (x \leq y \supset \exists z (x \leq z \wedge z \leq y))$$

But we can combine the insight (ii) with the older idea that quantifiers take two arguments, as long as we allow these arguments to be open formulas, not terms.[2]

2.1.2 Semantics of binary quantifiers

The semantics of binary quantifiers is relatively straightforward. For comparison, here is how we define truth in a model on an assignment for formulas headed by a unary quantifier:

$\vDash^{v}_{\mathcal{M}} \forall \alpha \psi$ iff for *every* assignment v' such that $v' \sim_{\alpha} v$, $\vDash^{v'}_{\mathcal{M}} \psi$.

$\vDash^{v}_{\mathcal{M}} \exists \alpha \psi$ iff for *some* assignment v' such that $v' \sim_{\alpha} v$, $\vDash^{v'}_{\mathcal{M}} \psi$,

where we abbreviate

$v' \sim_{\alpha} v \Leftrightarrow v'$ agrees with v on the values of every variable except possibly α.

And here's how we'd do it for some binary quantifiers:

$\vDash^{v}_{\mathcal{M}} \mathit{all}_{\alpha}(\phi, \psi)$ iff for *every* assignment v' such that $v' \sim_{\alpha} v$ and $\vDash^{v'}_{\mathcal{M}} \phi$, $\vDash^{v'}_{\mathcal{M}} \psi$.

$\vDash^{v}_{\mathcal{M}} \mathit{some}_{\alpha}(\phi, \psi)$ iff for *some* assignment v' such that $v' \sim_{\alpha} v$ and $\vDash^{v'}_{\mathcal{M}} \phi$, $\vDash^{v'}_{\mathcal{M}} \psi$.

$\vDash^{v}_{\mathcal{M}} \mathit{most}_{\alpha}(\phi, \psi)$ iff for *most* assignments v' such that $v' \sim_{\alpha} v$ and $\vDash^{v'}_{\mathcal{M}} \phi$, $\vDash^{v'}_{\mathcal{M}} \psi$.

$\vDash^{v}_{\mathcal{M}} \mathit{at\text{-}least\text{-}two}_{\alpha}(\phi, \psi)$ iff for *at least two* assignments v' such that $v' \sim_{\alpha} v$ and $\vDash^{v'}_{\mathcal{M}} \phi$, $\vDash^{v'}_{\mathcal{M}} \psi$.

2.1.3 Most: an essentially binary quantifier

Frege noticed that one could *define* the traditional binary quantifiers used in syllogistic logic in terms of unary quantifiers and truth-functional connectives. You already know how that can be done:

$$\begin{aligned}
\mathit{some}_{\alpha}(\phi, \psi) &\Leftrightarrow \exists x (\phi x \wedge \psi x) \\
\mathit{all}_{\alpha}(\phi, \psi) &\Leftrightarrow \forall x (\phi x \supset \psi x) \\
\mathit{at\text{-}least\text{-}two}_{\alpha}(\phi, \psi) &\Leftrightarrow \exists x \exists y (x \neq y \wedge \phi x \wedge \phi y \wedge \psi x \wedge \psi y) \\
\mathit{at\text{-}most\text{-}one}_{\alpha}(\phi, \psi) &\Leftrightarrow \forall x \forall y ((\phi x \wedge \phi y \wedge \psi x \wedge \psi y) \supset x = y)
\end{aligned}$$

[2] The first use of binary quantifiers in modern logic was in Lindström 1966.

Exercise 2.1: Infinite domains

⋆There's no way to *say* "there are infinitely many *F*s" in standard first-order logic. Still, one can write sentences of first-order logic that only have models with infinite domains. Can you come up with one?

Our success here might encourage us to think that this trick can always be pulled off: given *any* binary quantifier, we can define it in terms of truth-functional connectives and unary quantifiers. But it turns out that this is not the case. And the problem is not just that some binary quantifiers (like 'a few' and 'enough') are vague and context-sensitive. There are perfectly precise binary quantifiers that cannot be defined in terms of unary quantifiers.

A paradigm example is 'most', interpreted as meaning *more than half*. You might think, initially, that the binary quantifier 'most' could be defined in terms of a unary quantifier '$\mathcal{M}$', where $\ulcorner \mathcal{M}x\phi \urcorner$ is true in a model just in case more objects in the domain satisfy ϕ than do not. But how? We might start by formalizing 'Most cows eat grass' as '$\mathcal{M}x(Cx \supset Gx)$', but this will be true in *any* model where cows make up fewer than half the objects in the domain, no matter how many of them eat grass. On the other hand, '$\mathcal{M}x(Cx \wedge Gx)$' will be true *only* in models where cows are the majority of objects in the domain. So neither definition captures the meaning of 'Most cows eat grass'. Of course, there are other things we could try. (Try them on your own, and convince yourself that nothing like this is going to work.[3])

2.1.4 Unary quantifiers beyond ∀ and ∃

We don't need to look to binary quantifiers to find quantifiers that resist definition in terms of '∃' and '∀'. Try defining '$\mathcal{M}$' (our unary quantifier "most objects in the domain") in terms of '∃' and '∀'. There are other unary quantifiers that cannot be defined in terms of '∃' and '∀' and '=', including 'there are finitely many', 'there are infinitely many', and 'there are an even number of'. Adding these quantifiers to standard first-order logic yields more expressively powerful logics.[4]

[3] This was proved by Barwise and Cooper (1981, Appendix C, C12 and C13).

[4] The bible for this kind of thing is Barwise and Feferman 1985. The original generalization of quantifiers, and the observation that there were unary quantifiers that could not be defined in terms of '∀', '∃', and '=', is due to Mostowski (1957).

2.1.5 Generalized quantifiers

Logicians and linguists have tried to generalize the notion of a quantifier in a precise way, in the theory of *generalized quantifiers*. The basic idea is that an *n*-ary quantifier Q expresses a quantitative relation among n sets and the domain.[5]

Let's think this through with some examples (where D is the domain):[6]

Quantifier	Condition expressed
$\forall x\phi$	$D \subseteq \{x : \phi x\}$
$\exists x\phi$	$\{x : \phi x\} \cap D \neq \emptyset$
$all_x(\phi, \psi)$	$\{x : \phi x\} \subseteq \{x : \psi x\}$
$most_x(\phi, \psi)$	$\lvert\{x : \phi x\} \cap \{x : \psi x\}\rvert > \lvert\{x : \phi x\} - \{x : \psi x\}\rvert$

2.2 Definite descriptions

A *definite description* is a phrase that purports to denote an object as the unique thing satisfying a certain description: for example, 'the present king of France', 'the first dog born at sea', 'the bed', 'Claire's birthday' (equivalent to 'the birthday of Claire'), and '2 + 6' (equivalent to 'the sum of 2 and 6'). Although not all definite descriptions have the form 'the ϕ', they can all be rephrased that way, so we'll talk in what follows as if all definite descriptions have that form.[7]

2.2.1 Terms or quantifiers?

It is natural to suppose that definite descriptions are singular terms. Like pronouns and proper names, definite descriptions are used to denote a specific object. Those who are impressed by this parallel take definite descriptions to be *terms* grammatically, and *referring expressions* semantically. But it is also possible to think of definite descriptions as quantificational, treating 'the' as a binary quantifier like 'all' and 'at most two'.

[5] A *quantitative* (or *topic-neutral*) relation among sets is one that does not depend on which particular individuals belong to the sets, but only on their relative quantities. This notion can be defined precisely in terms of invariance under permutations of the domain (for an introductory explanation, see MacFarlane 2017, §5). The point of this restriction is to rule out, for example, a unary quantifier meaning "all mammals" or "everything but the Eiffel Tower."

[6] See Appendix B if you are unfamiliar with the set-theoretic notation used here.

[7] You may be puzzled how 'the bed' could purport to denote the unique bed, when we all know that there are many beds. Such definite descriptions are sometimes called *incomplete*. If we think of definite descriptions as quantifiers, it is natural to think that in such cases the domain of quantification is implicitly restricted (say, to the furniture in a single room). This kind of restriction can be seen in other quantifiers as well: when you ask 'Does anybody know when the game starts?' you are not asking whether anyone in the world knows this, but whether anyone in some relevant group knows.

We said that a binary quantifier can be thought of as expressing a relation among sets.[8] For example, '*at-least-two*$_x(Fx, Gx)$' says that $|F \cap G| \geq 2$ (the set of elements common to F and G has two or more members), and '*most*$_x(Fx, Gx)$' says that $|F \cap G| > |F - G|$ (the set of elements common to F and G has more members than the set of elements that belong to F but not G). Can we give a parallel treatment of 'the'? Let us ask what must be the case for

(1) The F is G

to be true?

- Surely, there must *be* an F.
- And presumably there can't be more than one F.
- Finally, this F must be G.

Taken together, these conditions are plausibly necessary and sufficient for the truth of (1). But the combination of these conditions can be represented as a simple quantitative condition on sets:

(2) *the*$_x(Fx, Gx)$ iff $|F| = |F \cap G| = 1$.

We can define truth in a model on an assignment for our new quantifier as follows.

$\vDash^{v}_{\mathcal{M}}$ *the*$_\alpha(\phi, \psi)$ iff

i) there is exactly one assignment v' such that $v' \sim_\alpha v$ and $\vDash^{v'}_{\mathcal{M}} \phi$, and

ii) there is exactly one assignment v' such that $v' \sim_\alpha v$ and $\vDash^{v'}_{\mathcal{M}} \phi$ and $\vDash^{v'}_{\mathcal{M}} \psi$.

So is the meaning of 'the' in English accurately modeled by the quantifier *the*? As I mentioned, this is a contentious question. We don't use 'the F' when there is more than one (salient) F, or when there aren't any. So it is tempting to suppose that 'the F is G' just *means* that there is a unique (salient) F and it is G. If that's right, then 'the' in English is a quantifier. On the other hand, the quantificational analysis also predicts that a sentence like 'The present king of France is bald' should come out *false*. And that has seemed odd to many philosophers. Surely, they say, if there is no present king of France, then 'The present king of France is bald' fails to make a determinate claim—and so fails to be either true *or* false. One who uses this sentence to make an assertion may *presuppose* that there is a present king of

[8] In what follows, I will uses 'F' as an abbreviation for '$\{x : Fx\}$' when it is clear from context that a set is intended.

France, but it seems odd to say (as the quantificational account does) that the sentence *entails* this.[9]

2.2.2 Definite descriptions and scope

If definite descriptions are quantifiers, then they have *scopes*. This means that certain English sentences will be predicted to have two readings, depending on how the scope ambiguity is resolved. Consider, for example:

(3) Fifteen presidential candidates are not campaigning in California.

This might mean either of two things:

(3w) $\mathit{fifteen}_x(Px, \neg Cx)$
There are fifteen presidential candidates who are not campaigning in California.

(3n) $\neg\mathit{fifteen}_x(Px, Cx)$
It's not the case that there are fifteen presidential candidates who are campaigning in California. (There are only six.)

In (3w), the quantifier takes *wide scope* over the negation. In (3n), it takes *narrow scope* (and the negation takes wide scope).

Do we see this phenomenon with definite descriptions? Consider:

(4) The present king of France is not washing my car.

(4w) $\mathit{the}_x(Kx, \neg Wx)$
The present king of France is such that: he is not washing my car.

(4n) $\neg\mathit{the}_x(Kx, Wx)$
It's not the case that the present king of France is washing my car.

On the quantificational reading, (4w), but not (4n), entails that there *is* a present king of France. Can we use (4) to mean both (4w) or (4n)? Or can we get only one reading? Ask yourself whether (4) can be *true* if (as is actually the case) there is no present king of France.

2.2.3 Russell's theory of descriptions

Bertrand Russell was the first to give a quantificational account of definite descriptions (Russell 1905). However, Russell did not have the theory of generalized quantifiers at his disposal, so he did not analyze 'the' as a binary quantifier. Instead,

[9] If you'd like to explore this debate, see Ostertag 1998 for the classic papers, and Neale 1990 for an influential defense of the quantificational view.

he represented 'the F' as a kind of *term*, which he then showed how to eliminate in favor of (standard) quantifiers.

Russell's definite description terms are constructed using an upside-down iota (ι). ι is a variable-binding operator, just like '$\forall$' and '$\exists$', but unlike them it forms a *term*, not a *formula*.[10] If ϕ is a formula and α is a variable, then $\ulcorner \iota\alpha\phi \urcorner$ is a term. For example:

(5) the F
$\iota x Fx$

(6) the F that Gs b
$\iota x(Fx \wedge Gxb)$

Terms formed using ι can occur in formulas wherever other kinds of terms (variables and individual constants) can occur. For example:

(7) the F is H
$H\iota x Fx$

(8) the F Gs the H that Gs the K
$G(\iota x Fx)(\iota x(Hx \wedge Gx\iota y Ky))$

(Put parentheses around iota-terms when there is threat of ambiguity.)

Russell understood the terms formed using his upside-down iota not as genuine terms, but as "incomplete symbols." In effect, he took formulas containing iota-terms to be *abbreviations* for formulas not containing them.

It is not hard to convince yourself that, unlike '$most_x$', the binary quantifier 'the_x' *can* be defined using standard unary quantifiers and identity:

(R1) $the_x(\phi x, \psi x) \equiv \exists x(\phi x \wedge \forall y(\phi y \supset y{=}x) \wedge \psi x)$
(The ϕ is ψ iff there is a unique ϕ and it is ψ.)

This is essentially the equivalence Russell uses to eliminate definite descriptions, but there is a twist due to his use of ι terms rather than binary quantifiers. As a first attempt at translating (R1) to Russell's notation, we might try:

(R2) $\psi\iota x\phi x \equiv \exists x(\phi x \wedge \forall y(\phi y \supset y{=}x) \wedge \psi x)$.

However, there is a problem with (R2) as it stands. Consider the formula

(9) $\neg R\iota x Px$.

Taking ψx to be 'Rx', (R2) says that (9) is equivalent to

(10) $\neg\exists x(Px \wedge \forall y(Py \supset y{=}x) \wedge Rx)$.

[10] This sort of operator is sometimes called a *subnector*.

But, taking ψx to be '$\neg Rx$', (R2) says that (9) is equivalent to

(11) $\exists x(Px \wedge \forall y(Py \supset y{=}x) \wedge \neg Rx)$.

Clearly (9) can't be equivalent to both (10) and (11), because they aren't equivalent to each other! So our rule (R2) is not sound.

What we need to solve this problem is a way of indicating the scope of definite descriptions written using iota-terms. Russell and Alfred North Whitehead do this in *Principia Mathematica* (Russell and Whitehead 1910) by putting a copy of the iota term in square brackets in front of the description's scope.[11] So, the narrow-scope reading of (9) would be written

(12) $\neg[\imath x Px]R\imath x Px$

and the wide-scope reading would be written

(13) $[\imath x Px]\neg R\imath x Px$.

(Note that the bracketed iota-term serves no function other than to indicate scope.)

Using this notation, we can write a (sound) version of our equivalence rule:

(R3) $[\imath x\phi x]\psi\imath x\phi x \equiv \exists x(\phi x \wedge \forall y(\phi y \supset y{=}x) \wedge \psi x)$.

Following Russell and Whitehead, we will adopt the convention that if the scope-indicator is omitted, the iota-term will be assumed to have the narrowest possible scope. Thus,

(14) $\neg R\imath x Px$

is to be read as

(15) $\neg[\imath x Px]R\imath x Px$,

which according to (R3) is equivalent to

(16) $\neg\exists x(Px \wedge \forall y(Py \supset y{=}x) \wedge Rx)$.

2.2.4 Proofs

Since for any formula containing the quantifier '*the*' or Russell's '$\imath$' operator we can always find an equivalent formula that uses only the standard quantifiers, it is easy to extend our proof system to accommodate definite descriptions.

[11] For a more detailed exposition, see Neale 2001, pp. 95–6.

RE (Russellian Equivalences)

Whenever a formula contains the right-hand side of an instance of (R1) or (R3), you may replace it with the left-hand side, and vice versa, with justification "RE." This is a substitution rule, so it may be used even on subformulas. Examples:

1	$the_x((Fx \wedge Gx), Hx)$	Hyp
2	$\exists x((Fx \wedge Gx) \wedge \forall y((Fy \wedge Gy) \supset y{=}x) \wedge Hx)$	RE 1 ($\phi x : Fx \wedge Gx, \psi x : Hx$)

(2.1)

1	$\exists x(Gx \wedge \forall y(Gy \supset y{=}x) \wedge (Fx \supset Hx))$	Hyp
2	$[\imath xGx](F\imath xGx \supset H\imath xGx)$	RE 1 ($\phi x : Gx, \psi x : Fx \supset Hx$)

(2.2)

1	$Ga \supset the_x(Fx, Gx)$	Hyp
2	$Ga \supset \exists x(Fx \wedge \forall y(Fy \supset y{=}x) \wedge Gx)$	RE 1 ($\phi x : Fx, \psi x : Gx$)

(2.3)

Notes:

1. Although terms formed using '$\imath$' are grammatically terms, they do not function *semantically* as terms (on Russell's account). Thus, in specifying a model, you do *not* specify an interpretation for these terms.

2. You cannot use these terms to get substitution instances when doing ∀ Elim, ∀ Intro, ∃ Elim, or ∃ Intro. For example, you cannot instantiate '$\forall x(x{=}x)$' with '$\imath xFx$' to get '$\imath xFx{=}\imath xFx$'. You'd better not be able to, because '$\imath xFx{=}\imath xFx$' implies '$\exists xFx$'. So you'd be able to prove the existence of an F for any F!

2.3 Second-order quantifiers

Recommended reading

George Boolos, "To Be is to Be a Value of a Variable (or to Be Some Values of Some Variables)" (Boolos 1984).

Exercise 2.2: Definite descriptions

1. How would you express the following sentences in logical notation? Do it first using the generalized quantifier *the*, and then using the Russellian *ɪ* operator.
   a) The man who killed Kennedy is a murderer.
   b) The shortest spy is the tallest pilot.
   c) Not every woman likes her (own) father.
2. Give a deduction of '$\exists xFx$' from '$[\imath xFx](\imath xFx{=}\imath xFx)$'.
3. Show that the = Elim rule is still valid when one term has the form '$\imath xFx$', not just when both terms are individual constants. That is, give a deduction that shows the validity of the following:

$$\frac{a{=}\imath xFx \quad Ga}{G\imath xFx}$$

4. As noted above, you cannot use an iota-term to instantiate a quantifier in the quantifier introduction and elimination rules. But surely 'something is a gas giant' follows from 'the farthest planet is a gas giant'. Show how you can get this conclusion using our rules.

So far, we have only considered quantifiers whose variables occur in places that could be occupied by terms. We'll say that a variable is in *term position* when the result of substituting an individual constant for (free occurrences of) the variable would be grammatical. *First-order logic* uses only quantifiers that bind variables in term position. *Second-order logic* includes quantifiers that bind variables in predicate position. Thus,

(17) $\exists XXa$

is a formula of second-order, but not first-order logic, because the variable 'X' occurs in a position that could be occupied by a one-place (or *monadic*) predicate. Similarly, in

(18) $\exists W\forall yWyy$

the variable 'W' occurs in a position that could be occupied by a two-place predicate.

This is, of course, only the tip of the iceberg. In principle, quantifiers might bind variables in other grammatical categories. For example, in

(19) $\exists p(p \supset \bot)$

the bound variable occupies *sentence position*; it occurs in a place where a sentence or formula could go. We can even imagine a quantifier that binds variables in *binary connective position*!

(20) $\exists * ((A * B) \equiv (B * A))$

We will return to quantification into sentence position in §2.4. In this section, we will mainly be concerned with second-order quantifiers, and specifically with quantifiers into monadic predicate position. We will mainly be concerned with two questions. First, the question of *expressive power*: Are there thoughts we can express with second-order quantifiers but not with first-order quantifiers? Second, the question of *meaning*: What, exactly, do the second-order quantifiers mean?

2.3.1 Standard semantics for monadic second-order logic

Although we will shortly be considering an alternative way of understanding second-order quantifiers, it will be helpful to begin with the standard treatment, building on our treatment of standard first-order logic in §1.2.

In our grammar, we will need to add a stock of second-order variables:

- A *second-order variable* is an uppercase W, X, Y, or Z, possibly with a numerical subscript (X_1, Y_{14}, etc.).

Second-order variables can all occupy the position of a monadic predicate, so we will need a new clause for atomic formulas:

- An *atomic formula* is an n-place predicate followed by n terms (for example, Fxy, G_1a_{15}) or a second-order variable followed by a term (for example, $W_{15}x$, Ya).

Our models will be just as in first-order logic: a domain of objects and an interpretation function. However, our assignment function v will now assign values to both first-order variables and second-order variables. First-order variables will be assigned objects from the domain, as before, and second-order variables will be assigned *subsets* of the domain. (Remember that the empty set is a subset of every domain, so it is always a possible assignment for a second-order variable.)

We will need a clause to handle atomic sentences in which the predicate place is occupied by a second-order variable:

- If ϕ is an atomic formula $\Omega\alpha$, where Ω is a second-order variable and α is a term, $\vDash^{v}_{\mathcal{M}} \phi$ iff $[\![\alpha]\!]^{v}_{\mathcal{M}} \in v(\Omega)$.

Finally, we will need clauses for the second-order quantifiers:

- $\vDash^{v}_{\mathcal{M}} \forall\Omega\psi$ iff for *every* assignment v' such that $v' \sim_{\Omega} v$, $\vDash^{v'}_{\mathcal{M}} \psi$.
- $\vDash^{v}_{\mathcal{M}} \exists\Omega\psi$ iff for *at least one* assignment v' such that $v' \sim_{\Omega} v$, $\vDash^{v'}_{\mathcal{M}} \psi$.

The basic idea is that the second-order quantifiers range over all subsets of the domain of the model. Thus, for example,

(21) $\exists X(Xa \wedge \forall y(Xy \supset Fy)$

is true in a model D, I just in case there is some subset of D that contains $I(\text{`}a\text{'})$ and is a subset of $I(\text{`}F\text{'})$. Test your understanding by thinking of a model in which (21) is true and one in which it is false.

2.3.2 Expressive limitations of first-order logic

Are there thoughts we can express in second-order logic but not in first-order logic?

One way to get at this question is to ask whether the models of a second-order formula can always be picked out as the models of a first-order formula. If the answer is yes, it means that second-order formulas do not give us any representational power beyond what first-order formulas provide. If the answer is no, it means that second-order formulas do allow us to represent scenarios that can't be captured using first-order formulas.

One might think it obvious that the answer is yes. Consider, for example,

(22) $\exists X(Xa \wedge \neg Xb)$

This says that there is a subset of the domain of the model that contains the denotation of '*a*' but not the denotation of '*b*'. How can we say this in first-order logic, which doesn't have quantifiers that range over subsets of the domain?

It turns out, though, that the same class of models can be represented by the first-order sentence

(23) $\neg(a{=}b)$

In any model where (23) is true, '*a*' and '*b*' will have different denotations, so there will be a subset of the domain to which the denotation of '*a*' but not the denotation of '*b*' belongs. Conversely, in any model where there is a subset of the domain to which the denotation of '*a*' but not the denotation of '*b*' belongs, (23) will be true. So (22) and (23) are made true by the same models.

Another way to get at the question is to ask whether there are English sentences that can be rendered in second-order logic but not first-order logic. Consider

(24) There are some critics who admire only one another (Boolos 1984).

This sentence doesn't seem to be quantifying over subsets of the domain, any more than

(25) There are some critics who admire only writers,

which has the straightforward first-order rendering

(26) $\exists x(Cx \wedge \forall y(Axy \supset Wy))$.

But, surprisingly, the meaning of (24) cannot be captured using first-order quantifiers.

Let's think about what (24) says. It says that there are some critics—call them the "In Group"—such that none of the members of the In Group admires herself or anyone outside of the In Group. To get clearer about it, try drawing some models in which it is true, and some in which it is false. Your models can consist of circles (critics), squares (non-critics), and arrows indicating who admires whom (see for example Fig. 2.1). In each model that makes (24) true, you should be able to answer the question: "Which are the critics who admire only one another?"

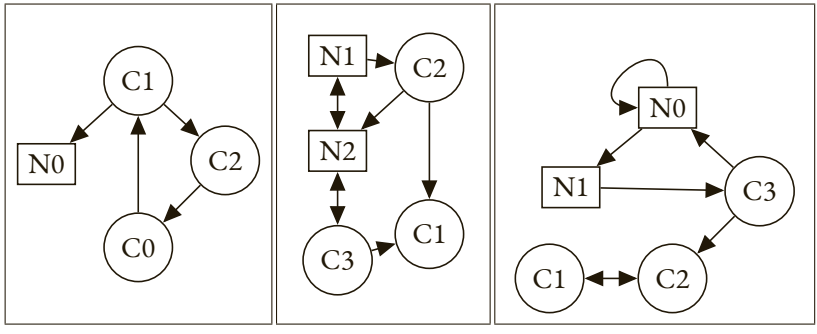


Figure 2.1: Which models make (24) true? Circles are critics, squares are noncritics, and the arrow is the admiring relation.

It turns out that there is no first-order sentence that is true in just those models in which there are some *C*s who admire only one another. The proof requires some metalogic, so we have placed it in Appendix C, for more adventurous readers. But you can get some insight by trying to formalize (24) in first-order logic, and seeing how you fail. This second-order representation gets the truth conditions right:

(27) $\exists X(\exists y Xy \wedge \forall y(Xy \supset Cy) \wedge \forall y \forall z[(Xy \wedge Ayz) \supset (y \neq z \wedge Xz)])$

Think of 'X' here as applying to the members of the In Group. '$\exists y Xy$' ensures that this group is nonempty, '$\forall y(Xy \supset Cy)$' that it consists only of critics, and '$\forall y \forall z[(Xy \wedge Ayz) \supset (y \neq z \wedge Xz)]$' that its members admire only other members of the group.

We have already seen (with 22 and 23) that some sentences that have natural second-order representations also have first-order equivalents. Boolos offers the example

(28) There are some monuments in Italy of which no one tourist has seen all.

Although this sentence is most naturally formalized as

(29) $\exists X(\exists x Xx \wedge \forall x(Xx \supset Mx) \wedge \neg \exists y(Ty \wedge \forall x(Xx \supset Syx)))$,

a bit of reflection shows that it is true just in case (a) there are some monuments in Italy, and (b) no tourist has seen all the monuments in Italy. And this can be expressed in first-order logic:

(30) $\exists x Mx \wedge \neg \exists x(Tx \wedge \forall y(My \supset Sxy))$.

Quine thought that

(31) Some of Fiorecchio's men entered the building unaccompanied by anyone else.

was an example of this kind. The natural symbolization is second-order: 'anyone else' means 'anyone else but *them*.'

(32) $\exists X(\exists x Xx \wedge \forall x(Xx \supset Fx) \wedge \forall x(Xx \supset Ex) \wedge \forall x \forall y((Xx \wedge Axy) \supset Xy))$.

Quine thought it could also be given a first-order rendering:

(33) $\exists x(Fx \wedge Ex \wedge \forall y(Axy \supset Fy))$.

But as Boolos shows, (33) can't be equivalent to (32), which (like 27) can be proved to have no first-order equivalent. (Can you see where Quine went astray?)

So far we've been representing sentences using a second-order existential quantifier. We can also find cases that need a universal quantifier. Often these will combine 'some' and 'if'. Compare:

(34) a. If there is a horse who can beat Cody, ...it ...
$\forall x((Hx \wedge Bxc) \supset \ldots x \ldots)$

b. If there are some critics who admire only each other, ...they ...
$\forall X([\exists y Xy \wedge \forall y(Xy \supset Cy) \wedge \forall y \forall z((Xy \wedge Ayz) \supset (y \neq z \wedge Xz))] \supset \ldots X \ldots)$.

A notable example of such a sentence is the principle of *mathematical induction*:

(35) If there are some things of which 0 is one, and of which every successor of one of them is one, then every number is one of them.
$\forall X([X0 \wedge \forall x \forall y((Xx \wedge Syx) \supset Xy)] \supset \forall x(Nx \supset Xx))$

This principle has no first-order equivalent. When we formalize arithmetic in first-order logic, we use instead an axiom *schema*, and posit as axioms all of its substitution instances:

(36) $[\phi 0 \wedge \forall x \forall y((\phi x \wedge Syx) \supset \phi y)] \supset \forall x(Nx \supset \phi x)$.

But the infinitely many instances of (36) are weaker than the second-order principle (35); they affirm that induction works for properties that can be defined with first-order formulas (the things that can be substituted for 'ϕ'), whereas (35) says that it works for *all* properties.

2.3.3 Set theory in sheep's clothing?

We've managed to formalize some sentences of English using second-order quantifiers, and we've seen that in many cases that is the *only* way we can formalize them. Let's now turn to the question of how, exactly, these second-order quantifiers are to be understood. For simplicity, let's focus on the simplest second-order sentence:

(37) $\exists X Xa$

We might understand (37) to be saying one of the following:

(38) a. There is a property that *a* possesses.
b. There is a concept that applies to *a*.
c. There is a set of which *a* is a member.

But if we understand the second-order quantifier in one of these ways, we would seem to be vulnerable to Quine's objection:

> To put the predicate letter 'F' in a quantifier...is to treat predicate positions suddenly as name positions, and hence to treat predicates as names of entities of some sort. (Quine 1970, pp. 66–7)

Quine's point is that these renderings would be appropriate for first-order sentences in which a variable ranging over properties, concepts, or sets occurs in name position:

(39) a. $\exists x Hax$
b. $\exists x Axa$

Exercise 2.3: Second-order quantifiers

1. Symbolize the following English sentences in second-order logical notation:

   a) Some cats and some dogs came to Harry's Bar, and each of the cats danced with at least one of the dogs.

   b) There are some propositions such that the negation of any of them is one of them.

   c) If there are some people on the street, at least one of them will get shot.

   d) If some numbers are such that each of them is the product of two of them, then there are more than two of them.

   e) Some critics admire only writers who hate most of them.[a]

2. For which of the above sentences can you find equivalent first-order formulas? (State the equivalent formulas, or 'can't find one.')

3. ⋆Give an argument for the equivalence of (a) and (Z) from Boolos 1984, p. 439:

   $\exists x(\exists y Azy \wedge \forall x((z{=}x \vee Azx) \supset \forall y(Axy \equiv ((z{=}y \vee Azy) \wedge y \neq x))))$ (a)

   $\exists X(\exists x \exists y(Xx \wedge Xy \wedge x \neq y) \wedge \forall x(Xx \supset \forall y(Axy \equiv (Xy \wedge y \neq x))))$ (Z)

4. ⋆Can you think of a situation in which (32) would be true but (33) false, under the intended interpretation (where 'Fx' means 'x is one of Fiorecchio's men', 'Ex' means 'x entered the building', etc.)? If not, how can we make sense of Boolos's claim that the sentences are not equivalent?

5. ⋆How would you symbolize 'Some teachers moved the piano across the room'?

[a] In this exercise, you may use a dyadic first-order quantifier '$most_x$' to symbolize 'most'.

c. $\exists x\ a \in x$

Using the notation of second-order quantification to express these thoughts, Quine thinks, is just a confusion, and one that engenders a dangerous conflation of predicates with names. Second-order logic, he concludes, is just "set theory in sheep's clothing" (Quine 1970, p. 68).

Moreover, when we say

(40) There are some critics who admire only each other,

it doesn't seem as if we're talking about special entities like sets or properties or concepts. We're just talking about critics. It seems perfectly coherent to say, for example,

(41) There are some critics who admire only each other, and there are no sets.

As Boolos puts it, "...there may be a set containing all trucks, but that there is certainly doesn't seem to *follow* from the truth of 'There are some trucks of which every truck is one' " (Boolos 1984, p. 447). Indeed, this last sentence seems to say little more than that there are trucks.

Indeed, as Boolos points out, there are some things we'd like to be able to say about sets using our second-order language that we *couldn't* coherently say if our second-order quantifiers were really quantifiers over sets. For example,

(42) $\exists X(\exists x Xx \wedge \forall x(Xx \supset Sx) \wedge \forall x((Sx \wedge x \notin x) \supset Xx))$
There are some sets of which every set that is not a member of itself is one.

Boolos thinks that (42) is equivalent to the first-order sentence

(43) $\exists x(Sx \wedge \forall y((Sy \wedge y \notin y) \supset Sy))$
There are some sets and every set that is not a member of itself is a set,

which is true. But if we take (42) to be a disguised set-theoretic statement, namely

(44) There is a set of sets of which every set that is not a member of itself is a member,

we get something that is *false* according to standard Zermelo-Fraenkel (ZF) set theory, for reasons relating to Russell's paradox. (If there is such a set, is it a member of itself, or not?)

2.3.4 Boolos's plural interpretation

Boolos's alternative suggestion is that the second-order quantifiers don't range over anything other than the objects the first-order quantifiers range over. It's just that they range over them *plurally* instead of singly:

> We need not construe second-order quantifiers as ranging over anything other than the objects over which our first-order quantifiers range, and in the absence of other reasons for thinking so, we need not think that there are collections of (say) Cheerios, in addition to the Cheerios. Ontological commitment is carried by our *first*-order quantifiers; a second-order quantifier needn't be taken to be a kind of first-order quantifier in disguise, having items of a special kind, collections, in its range. (Boolos 1984, p. 449)

To support this claim, Boolos provides a procedure for translating any sentence of (monadic) second-order logic into a sentence of English that does not make any mention of sets or collections, but instead uses normal plural constructions that we understand independently of set theory. The procedure is summarized in Table 2.1. The English sentences that his translations produce are stilted and slightly augmented by subscripts (though these could be eliminated in principle by phrases such as 'the former' and 'the latter'). But they are "proper sentences of English which, with a modicum of difficulty, can be understood and seen to say something true" (Boolos 1984, p. 442).

Let's try using this procedure to get a translation of

(45) $\exists X \forall x(Fx \equiv Xx)$.

The main operator is '$\exists X$', so we apply the rule for '$\exists V \psi$', with V='X' and ψ='$\forall x(Fx \equiv Xx)$':

(46) 'Either there are some things that$_X$ are such that ' $\frown$ Tr('$\forall x(Fx \equiv Xx$')) $\frown$ ' or '$\frown$ Tr('$\forall x(Fx \equiv x \neq x$')).

The last part (after 'or') is just standard first-order stuff. So we just need to compute Tr('$\forall x(Fx \equiv Xx$')). Applying the rule for '$\forall v$', we get:

(47) 'everything$_x$ is such that '$\frown$ Tr('$Fx \equiv Xx$').

Tr('$Fx \equiv Xx$') is

(48) Tr('Fx') $\frown$' if and only if '$\frown$ Tr('Xx').

Tr('Fx') is just 'it$_x$ is a frog,' and Tr('Xx') is 'it$_x$ is one of them$_X$.' Putting it all together then, we get:

(49) Either there are some things that$_X$ are such that everything$_x$ is such that it$_x$ is a frog if and only if it$_x$ is one of them$_X$, or everything is such that it$_x$ is a frog if and only if it$_x$ is not identical with itself.

Which is intelligible English, though not something you'd want to write! Once you have this, you can try to get a smoother version:

(50) Either there are some things such that a thing is one of them just in case it's a frog, or nothing is a frog.

ϕ	$\mathrm{Tr}(\phi)$
v (variable)	$\ulcorner \mathrm{it}_v \urcorner$
a (constant)	'Alex' [for example]
$F\alpha$ (predicate)	$\mathrm{Tr}(\alpha)$ ⌢' is a frog' [for example]
$V\alpha$ (predicate variable)	$\mathrm{Tr}(\alpha)$ ⌢ $\ulcorner$ is one of them$_V$ $\urcorner$
$\alpha{=}\beta$	$\mathrm{Tr}(\alpha)$ ⌢' is identical with '⌢ $\mathrm{Tr}(\beta)$
$\phi \wedge \psi$	$\mathrm{Tr}(\phi)$ ⌢' and '⌢ $\mathrm{Tr}(\psi)$
$\phi \vee \psi$	$\mathrm{Tr}(\phi)$ ⌢' or '⌢ $\mathrm{Tr}(\psi)$
$\phi \supset \psi$	$\mathrm{Tr}(\phi)$ ⌢' only if '⌢ $\mathrm{Tr}(\psi)$
$\phi \equiv \psi$	$\mathrm{Tr}(\phi)$ ⌢' if and only if '⌢ $\mathrm{Tr}(\psi)$
$\neg\phi$	'it is not the case that '⌢ $\mathrm{Tr}(\phi)$
$\exists v\psi$	$\ulcorner$there is something$_v$ such that $\urcorner$ ⌢ $\mathrm{Tr}(\psi)$
$\forall v\psi$	$\ulcorner$everything$_v$ is such that $\urcorner$ ⌢ $\mathrm{Tr}(\psi)$
$\exists V\psi$	'either there are some things that$_V$ are such that '⌢ $\mathrm{Tr}(\psi)$ ⌢' or '⌢ $\psi^\dagger$, where $\psi^\dagger$ is the result of substituting every occurrence of $\ulcorner V\alpha \urcorner$ in ψ with $\ulcorner \alpha{\neq}\alpha \urcorner$
$\forall V\psi$	$\mathrm{Tr}(\ulcorner \neg\exists V\neg\phi \urcorner)$

Table 2.1: Boolos's translation scheme. Note that v is used here as a metavariable ranging over first-order variables, and V as a metavariable ranging over second-order (predicate) variables.

The reason for the complicated clause for '∃' is that English 'some things' implies at least one thing, *but the second-order quantifier does not. So, for example, '$\exists X\forall x(Xx \equiv x{\neq}x)$' is true, but 'there are some things such that a thing is one of them if and only if it is not identical to itself' is not. $\psi^\dagger$ says, essentially, that ψ is true when nothing is X.*

2.3.5 Beyond monadic second-order logic

We have seen that monadic second-order logic has greater expressive power than first-order logic, and we have explored Boolos's interpretation of the monadic second-order quantifiers as plural quantifiers. We can get even more expressive

Exercise 2.4: Boolos's translation scheme

Translate the following second-order formulas into English, using Boolos's translation method. (You may use subscripted pronouns, like 'them$_X$,' but after giving a translation using these, try to give a more idiomatic version.) Take 'Fx' to mean 'x is a frog,' 'a' to mean 'Al' and 'b' to mean 'Bo'.

1. $\exists X(Xa \vee Xb)$
2. $\exists X(\forall x(Xx \supset Fx) \wedge \neg Xa)$
3. $\neg\forall X\forall x(Xx \supset Fx)$

power by moving to full second-order logic, which allows variables that can stand in place of *polyadic* predicates (predicates with two or more argument places).[12]

Consider, for example, the notion of *equinumerosity* (sameness of number). It is standard in mathematics to define equinumerosity in terms of one-to-one correspondence. To say that the knives and the forks on the dinner table are the same in number is to say that there is a way of relating them one-to-one, with each knife paired with a single fork, and no forks that are not paired with any knife. There is no way to define this notion in first-order logic—no way, that is, to fill out the right-hand side of this equivalence using just first-order vocabulary:

There are equally many Fs as Gs $\equiv$ ...

(Take a few minutes to try.) Adding monadic second-order quantification does not help. But with quantification over relation variables, we can pull it off. Define[13]

$$X \sim Y \equiv_{def} \exists R(\overbrace{\forall x(Xx \supset \exists y(Yy \wedge Rxy \wedge \forall z((Yz \wedge Rxz) \supset z{=}y)))}^{\text{each } X \text{ } R\text{s a unique } Y} \wedge \underbrace{\forall y(Yy \supset \exists x(Xx \wedge Rxy))}_{\text{each } Y \text{ is } R\text{d by an } X}) \tag{2.4}$$

Then we can express 'there are equally many Fs as Gs' as

$$\exists X\exists Y(\forall z(Xz \equiv Fz) \wedge \forall z(Yz \equiv Gz) \wedge X \sim Y)$$

[12] The modifications needed to the standard semantics we presented for monadic second-order logic in §2.3.1 are the obvious ones: for example, second-order two-place relation variables range over sets of ordered pairs of objects from the domain.

[13] Here '~' expresses a relation whose argument places are filled by monadic second-order variables. If we do not want to expand our grammar to allow such higher-order relations, we can simply regard it as a notational convenience which can be eliminated using the equivalence.

Once we have the notion of equinumerosity, we can use it to define finitude and infinitude. If there are finitely many Fs, then there is no one-to-one function relating the Fs to just *some* of the Fs (say, all the Fs but one). On the other hand, an infinite set can always be mapped one-to-one onto a proper part. For example, we can map the natural numbers to the natural numbers greater than 0 as follows:

$$\begin{array}{ccccccccc} 0 & 1 & 2 & 3 & 4 & \ldots & n & & \ldots \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow & & \\ 1 & 2 & 3 & 4 & 5 & \ldots & n+1 & & \ldots \end{array}$$

Take a minute to convince yourself that this is a one-to-one mapping in the sense defined above: to each number in the top row, we have paired a unique number in the bottom row, and for every number in the bottom row, there is a number in the top row that is paired with it. So, somewhat counterintuitively, the natural numbers including 0 are equinumerous with the natural numbers greater than 0! We can use this funny property as a characterization of infinitude. There are infinitely many Fs if there is a one-to-one mapping between the Fs and all the Fs but one:

$$\textit{infinite}\, x\, Fx \equiv$$
$$\exists X \exists Y \Big(\underbrace{\forall x(Xx \equiv Fx)}_{\text{the } X\text{s are the } F\text{s}} \wedge \underbrace{\exists z(Xz \wedge \forall x(Yx \equiv (Xx \wedge x \neq z)))}_{\text{the } Y\text{s are all the } X\text{s but one}} \wedge X \sim Y \Big)$$

And there are finitely many Fs if there are not infinitely many Fs:

$$\textit{finite}\, x\, Fx \equiv \neg\, \textit{infinite}\, x\, Fx$$

In addition to equinumerosity, finitude, and infinitude, many other mathematical concepts that are not definable in first-order logic can be defined in second-order logic (countability and well-foundedness are two other important examples). Indeed, full second-order logic is so expressive that many questions about validity in second-order logic rest on disputed questions in set theory, such as the truth of the Continuum Hypothesis.[14] This fact has been taken to impugn the claim of second-order logic to be logic, and to support Quine's view that it is "set theory in sheep's clothing."[15]

Note, also, that Boolos's plural interpretation is defined only for monadic second-order variables. It does not directly give us a way to interpret formulas

[14] The Continuum Hypothesis is the claim that every subset of the real numbers is equinumerous with either the natural numbers or the reals: in other words, there are no cardinal numbers between the number of the naturals and the number of the reals. Both the Continuum Hypothesis and its negation are consistent with the standard axioms for set theory.

[15] For useful critical examinations of Quine's charges, see Boolos 1975 and Shapiro 1991.

Exercise 2.5: Defining generalized quantifiers in second-order logic

We saw in §2.1.3 that the binary quantifier '$[most_x \phi](\psi)$' cannot be defined in terms of the standard first-order universal and existential quantifiers.

1. Define 'there are more *F*s than *G*s' using second-order logic. That is, produce a sentence of second-order logic that is true if and only if there are more things in the extension of '*F*' than in the extension of '*G*'.
2. Define 'most *F*s are *G*s' in second-order logic.
3. ⋆Define 'there are finitely many *F*s and there are twice as many *G*s as *F*s' in second-order logic.

such as (2.4) in terms of plural quantification. Boolos notes that we can simulate quantification over relations using plural quantification over ordered pairs (Boolos 1985, 330 n. 4). But this need for a special ontology of pairs arguably takes us outside the realm of pure logic.

2.4 Substitutional quantifiers

Recommended reading

Ruth Barcan Marcus, "Interpreting Quantification" (Marcus 1962).

2.4.1 Objectual and substitutional quantification

In giving the semantics of regular (*objectual*) quantifiers, we say which assignments of objects from the domain to the variables must make the embedded open formula true in order for the whole quantified formula to be true. We say, in effect: this formula (with a quantifier binding the variable α) is true just in case the open formula we get by stripping off the quantifier is true on all/some assignments of values to α.

Substitutional quantifiers have a different semantics. We say, in effect: this formula (with a quantifier binding α) is true just in case all/some of the formulas you'd get by stripping off the quantifier and replacing α with an individual constant (or *name*) are true. Here we do not talk of variable assignments, or truth on a

variable assignment, at all. Instead, we talk of substitution instances, and plain truth.[16]

It will often be useful to consider both kinds of quantifiers at the same time, so we'll explain the semantics of substitutional quantifiers in the framework of models and assignments we've been using all along. For the same reason, we'll use different symbols for the substitutional quantifiers: 'Σ' for the existential and 'Π' for the universal.

Substitutional quantifiers Σ and Π Where ϕ is a formula, α is a variable, and $\phi[\beta/\alpha]$ is the result of substituting β for every occurrence of α in ϕ,

$\vDash^{a}_{\mathcal{M}} \Sigma\alpha\phi$ iff for some individual constant β in the language, $\vDash^{a}_{\mathcal{M}} \phi[\beta/\alpha]$.

$\vDash^{a}_{\mathcal{M}} \Pi\alpha\phi$ iff for every individual constant β in the language, $\vDash^{a}_{\mathcal{M}} \phi[\beta/\alpha]$.

It is easy to describe a model on which '$\exists xFx$' is true but 'ΣxFx' is false. Take the domain to be all natural numbers, take the extension of 'F' to be the set of even numbers, and take all the individual constants to denote the number 1. (Verify for yourself that 'ΣxFx' is false on this interpretation.)

What is substitutional quantification good for? Several interesting applications have been explored in the philosophical literature.

2.4.2 Nonexistent objects

Consider the inference:

(51)
Pegasus is a winged horse

Something is a winged horse

Marcus (1962) notes that if we interpret the conclusion with an objectual quantifier, '$\exists x(Wx \wedge Hx)$', then it can only be true (on the intended interpretion of 'W' and 'H') if the domain contains at least one object that is a winged horse. And that seems to saddle us with an unpleasant dilemma. If we take the conclusion to be true, we must allow "nonexistent objects" into our domain.[17] We can avoid this ontological extravagance by saying that the conclusion isn't true. But then, because the argument is valid, we must deny that the premise is true. And isn't it true that Pegasus is a winged horse (and false that he is a three-legged cow)?

[16] This interpretation of the quantifiers was first proposed and defended by Ruth Barcan Marcus (Marcus 1962; Marcus 1972).

[17] This approach is often associated with the philosopher Alexius Meinong. Quine raises some difficulties for it in his classic article "On What There Is" (Quine 1948).

Marcus offers a way out of this dilemma: instead of interpreting the quantifier in the conclusion objectually, we can interpret it substitutionally. For

(52) $\Sigma x(Wx \wedge Hx)$

to be true, it is only required that an instance, such as

(53) $Wa \wedge Ha$

be true. No other demand is made on the domain. So we can accept both the premise and the conclusion of (51) without countenancing nonexistent objects.

You might object that the truth of (53) requires that there be an object in the domain for the constant '*a*' to denote, so we'll need Pegasus in our domain after all. This doesn't derail Marcus's approach altogether, but it shows that it can only succeed if we have an alternative account of the truth conditions of (at least some) atomic sentences, one that allows 'Pegasus is a winged horse' to be true even if 'Pegasus' does not denote anything.

2.4.3 Quantifying into attitude reports

Suppose we think that

(54) Caesar believed that Juno favored him.

It is natural to want to existentially generalize:

(55) There's someone who Caesar believed favored him. (Namely, Juno.)

But if 'someone' is an objectual quantifier, this requires that the domain contain Juno. And we might think that Juno doesn't exist.

Similar problems arise when people have different ways of thinking about an existing object. Consider:

(56) a. Sarah thinks Eminem is clever.
b. Sarah does not think Marshall Mathers is clever.

Though Sarah doesn't know this, in fact, Enimem *is* Marshall Mathers. So if it makes sense to quantify objectually inside the attitude report, it should be true that

(57) There is someone, *x* (= Eminem, a.k.a. Marshall Mathers) such that Sarah thinks *x* is clever and Sarah does not think *x* is clever.

But surely no element of our domain can have the contradictory properties of being thought clever by Sarah and *not* being thought clever by Sarah. What has gone wrong here?

It is commonly held that 'thinks that' creates a special context in which proper names do more than simply pick out their bearers.[18] In these contexts, it matters not just whom the name denotes, but which name is used. If that is right, then it doesn't make sense to do existential generalization on 'Eminem' in (56a) or 'Marshall Mathers' in (56b).

But it seems quite natural to quantify into these contexts. We might want to say, for example,

(58) There is someone who Sarah thinks is clever.

Substitutional quantification makes sense here, even if the truth values of the substitution instances ('Sarah thinks Ken is clever', 'Sarah thinks Eminem is clever', 'Sarah thinks Marshall Mathers is clever', etc.) depend on which name is used, and not just on the object it denotes.

2.4.4 Sentence quantifiers

So far, we have been considering substitutional interpretations of first-order quantifiers—quantifiers whose variables occur in places that could be occupied by terms. But one natural use for the substitutional interpretation would be to interpret quantification into *sentence position*, as in

(59) $\exists p(p \supset \bot)$

It is crucial here not to confuse *being in sentence position* with *ranging over sentences*. A quantifier that binds variables in name position can range over sentences. For example, if I say

(60) Every sentence I have written is convoluted,

I have quantified over sentences, but the quantifier I have used binds variables in name position:

(61) $\forall x((Sx \wedge Wix) \supset Cx)$

In (59), by contrast, the variable 'p' occurs in the grammatical position where a formula (or sentence) could go. (Putting a name before '$\supset \bot$' would yield something ungrammatical.)

It seems doubtful that we have quantifiers into sentence position in natural languages. If we did, we could use them to say things like

(62) For all p, if the senator says that p then $\neg p$.

Here it's clear that 'p' is in sentence position: only a formula can grammatically fit after a '$\neg$', or after 'says that'. What is the closest you can come to (62) in ordinary English?

[18] For the seminal discussion, see Frege 1892/1980.

2.4.5 Quantifying into quotes

Suppose you wanted to say,

(63) For some *x*, Athos asked *x*, 'Do you know where *x* is?'

The underlying thought seems intelligible. It is true if, for example, Athos asked Madame Bonacieux, 'Do you know where Madame Bonacieux is?' As it stands, though, there is a problem: the second '*x*' is in quotes, and Athos did not use variables in his speech.

We could try to fix *that* problem by using quasiquotation:

(64) For some *x*, Athos asked *x*, ⌜ Do you know where *x* is? ⌝.

But this isn't right either. For this sentence to be true on an assignment, the assignment must assign a *name* to '*x*'. That is required by the quasiquotes: only a linguistic expression can be concatenated. But if '*x*' denotes a *name*, then the sentence can't be true: Athos didn't ask a *name* to answer a question.

If we construe the quantifier substitutionally, we can make good sense of our sentence:

(65) Σ*x*(Athos asked *x*, 'Do you know where *x* is?')

(65) is true just in case there is a name in our language which, when put into the slots below, makes a true sentence:

(66) Athos asked __, 'Do you know where __ is?')

The fact that one of these slots is inside quotes doesn't matter in the least to the substitutional quantifier.

2.4.6 Defining truth

In his classic article "The Concept of Truth in Formalized Languages" (Tarski 1935; Tarski 1983b) Alfred Tarski noted that any good definition of truth (for sentences) ought to imply all instances of the schema

(67) '*S*' is true iff *S*

where *S* is a sentence. He then proceeded to show how to give such a definition, restricted to a particular language, using considerable ingenuity and some set theory. (His work on defining truth is the basis of the definitions of truth in a model we looked at in §1.2.3.)

Someone might wonder: isn't it just *trivial* to give a definition that meets Tarski's criteria?

(68) $\forall x(x \text{ is true} \equiv x)$

You can see the problem, I hope. If not, ask yourself: to what grammatical category does the bound variable '*x*' belong? Is it in name position or sentence position? (Be sure to look at *both* occurrences of '*x*'.)

What we need is clearly substitutional quantification:[19]

(69) $\Pi p(\text{`}p\text{'} \text{ is true} \equiv p)$

As Linsky (1972, p. 235) points out, this isn't general enough: it shows us how to deal with 'is true' when it applies to a quote name of a sentence, but not when it applies to some other term denoting a sentence, like 'the third sentence on this page'. The fix is easy enough, though:

(70) $\forall x(x \text{ is true} \equiv \Sigma p(x{=}\text{`}p\text{'} \wedge p))$

However, there are two rather serious problems for the project of defining truth this way.

2.4.7 Quantifying into quotes and paradox

The first is that substitutional quantification into quotes leads to paradox. The argument can be found in Tarski 1983b, pp. 161–2. Here is Tarski's presentation:

> Let the symbol '*c*' be a typographical abbreviation of the expression '*the sentence printed on this page, line 6 from the top*'. We consider the following statement:
>
> > *for all p, if c is identical with the sentence 'p', then not p* [Note: This is, in fact, the sentence printed on the sixth line from the top of the page in Tarski's article.]
>
> (if we accept [70] as a definition of truth, then the above statement asserts that *c* is not a true sentence).
>
> We establish empirically:
>
> (α) *the sentence 'for all p, if c is identical with the sentence 'p', then not p' is identical with c.*
>
> In addition we make only a single supplementary assumption which concerns the quotation-function and seems to raise no doubts:
>
> (β) *for all p and q, if the sentence 'p' is identical with the sentence 'q', then p if and only if q.*
>
> By means of elementary logical laws we can easily derive a contradiction from the premises (α) and (β).

[19] Here, following Tarski, we use '*p*' as a substitutional variable.

Let us try to reconstruct Tarski's reasoning in our notation. Let c denote the sentence

(c) $\Pi p(c=\text{`}p\text{'} \supset \neg p)$.

Then, Tarski claims, the following sentence should be true, since it just says what c is:

(α) $c=\text{`}\Pi p(c=\text{`}p\text{'} \supset \neg p)\text{'}$.

Suppose c is true. Then, instantiating 'p' in c with c itself, we get

(71) $c=\text{`}\Pi p(c=\text{`}p\text{'} \supset \neg p)\text{'} \supset \neg\Pi p(c=\text{`}p\text{'} \supset \neg p)$.

But the antecedent here is just (α), so by Modus Ponens we get

(72) $\neg\Pi p(c=\text{`}p\text{'} \supset \neg p)$.

This is just the negation of c. So we've shown that if c is true then c is false.

Now suppose c is false, that is,

(73) $\Sigma p(c=\text{`}p\text{'} \wedge p)$.

Then there is some q such that

(74) $c=\text{`}q\text{'} \wedge q$.

But then it follows from (α) that

(75) $\text{`}q\text{'}=\text{`}\Pi p(c=\text{`}p\text{'} \supset \neg p)\text{'}$.

By (β), we then have

(76) $q \equiv \Pi p(c=\text{`}p\text{'} \supset \neg p)$.

But from (74) we can get (by $\wedge$ Elim)

(77) q,

so we can conclude

(78) $\Pi p(c=\text{`}p\text{'} \supset \neg p)$,

which is of course c. So, if c is false, then it's true, and if it's true, then it's false. Paradox!

One way around this problem is to restrict the sentences that can be substituted for the sententential variable in (c). Kripke (1976, pp. 366–8) and Soames (1999, p. 87) suggest that we disallow substituends that themselves contain sentential variables. (Can you see how this would block Tarski's derivation?) Marcus (1972, p. 247) proposes limiting substituends for the sentence quantifiers to sentences containing fewer quantifiers than the original. Both approaches save us from paradox by weakening what is said by sentences containing substitutional quantifiers. One might ask, though, whether after this weakening (70) still captures the concept of truth.

2.4.8 The circularity worry

Even if the definition of truth using substitutional quantifiers can be made coherent, one might worry that it is somehow circular. Here's how the objection might go: If we interpret the sentence position quantifier in

(79) For some p, a = 'p' and p

substitutionally, then the sentence has the following sense: there is a sentence which, when it replaces the two occurrences of 'p' in the embedded formula, produces a *true* sentence. But our grasp of this sense presupposes an understanding of truth, so we cannot use this kind of substitutional quantifier to *define* truth—at least not if we hope to define truth in nonsemantic terms. As Mark Platts puts it: "The problem is clear: substitutional quantification is defined in terms of truth, and so cannot itself be used to define truth" (Platts 1979, pp. 14–15).

Soames (1999, p. 91) argues that this objection "embodies a subtle but fundamental mistake about the nature of substitutional quantification." He argues that although (79) is *true* if and only if there is a sentence which, when it replaces the two occurrences of 'p' in the embedded formula, produces a *true* sentence, it does not *mean* that there is a sentence which, when it replaces the two occurrences of 'p' in the embedded formula, produces a *true* sentence. According to Soames, "the proposition that is expressed by the [substitutional] quantification is not a metalinguistic proposition about expressions at all. To suppose otherwise is to confuse substitutional quantification with objectual quantification over expressions." He gives two arguments:

1. We define '$\wedge$' by saying that $\ulcorner A \wedge B \urcorner$ is true if and only if A is true and B is true. (This is the truth-table definition.) But this does not mean that 'Joe swims $\wedge$ Mary bikes' expresses a metalinguistic proposition about the truth of the sentences 'Joe swims' and 'Mary bikes'. So the fact that we explain the substitutional quantifier in terms of truth does not make it any less appropriate for use in a definition of truth than conjunction is.

2. The sentence 'Σn(n is hot)' does not express a metalinguistic proposition about sentences. For one can believe that for some n, n is hot without believing anything about sentences. Similarly, the proposition this sentence expresses would have been true even if the word 'hot' had meant something different, but the metalinguistic proposition might not have been true in that case.

However, even if one is persuaded that 'Σn(n is hot)' is not a metalinguistic claim about the truth of sentences, one might be sympathetic with Peter van Inwagen's complaint:

Exercise 2.6: Substitutional quantifiers

1. ⋆What introduction and elimination rules would work for the substitutional quantifiers?
2. How would you express what is said by (62) in ordinary English?
3. ⋆Which of the three approaches to the inference (51) in §2.4.2 do you think is best, and why? Write up your thoughts in a paragraph or two.
4. ⋆What do you think is the best response to van Inwagen's complaint about the intelligibility of substitutional quantification?

> If I could understand the sentence
>
> $S \quad \Sigma x(x \text{ is a dog})$
>
> then I could understand substitutional quantification. But I cannot understand this sentence. I cannot understand it because I do not know what proposition it expresses. (van Inwagen 1981)

Proponents of substitutional quantification say that S is true just in case some sentence that is the result of substituting a term for 'x' in 'x is a dog' is true. They deny that S *says* this. Van Inwagen's complaint is that they don't say what S *does* say.

The issues here are interesting and subtle, and it is worth thinking hard about Platts's objection, Soames's response, and van Inwagen's complaint.

Further readings

- On generalized quantifiers, the classic Barwise and Cooper 1981 is a very good read.
- On definite descriptions, see the anthology Ostertag 1998, which collects many classic papers.
- On second-order logic, see Shapiro 1991.
- On the interpretation of second-order quantifiers, see Quine 1970, ch. 5, Boolos 1975, and Rayo and Yablo 2001.
- On substitutional quantifiers, see Dunn and Belnap 1968, Linsky 1972, Marcus 1972, Kripke 1976, and van Inwagen 1981.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3 Modal Logic

Consider this argument:

(1)
Necessarily, if the door is open, Leo is outside.
It is possible that the door is open

It is possible that Leo is outside.

This looks like a valid argument. But we cannot evaluate it using the logical tools we have developed so far, because we have no way to represent sentence modifiers like 'necessarily' or 'it is possible that'. These modifiers are known as *modal operators*, and the branch of logic that concerns them is called *modal logic*.

In this chapter, we will learn the fundamentals of propositional modal logic. We will study a number of related modal logics, from both a semantic and a proof-theoretic point of view, and we will think about how different logics are appropriate for different interpretations of the modal operators. We will then consider some conceptual problems that arise when modal operators are combined with quantifiers, and consider how they might be resolved. Among these is the notorious *slingshot argument*, deployed by W. V. O. Quine to argue against quantified modal logic and by Donald Davidson to argue against a correspondence theory of truth. In assessing it we will draw on the work we have already done on definite descriptions.

3.1 Modal propositional logic

3.1.1 Grammar

In antiquity and the middle ages, words like 'might', 'must', 'possibly', and 'necessarily' were thought to modify the relation between a subject and predicate term in a proposition. In contemporary logic, by contrast, the modalities are understood as one-place sentential connectives. Just like '¬', these new connectives, '□' and '◇', can attach to any formula and yield a new formula.

Table 3.1: Interpretations of '□' and '◊'.

□	◊
it is logically necessary that	it is logically possible that
it could not have failed to happen that	it might have happened that
it must be the case that	it might be the case that
it is now settled that	it is still possible that
it is obligatory that	it is permitted that
it is provable that	it is not refutable that
A believes that	A's beliefs do not exclude that
A knows that	for all A knows, it may be true that

□ and ◊ If ϕ is a formula, then $\ulcorner \Box\phi \urcorner$ and $\ulcorner \Diamond\phi \urcorner$ are formulas.

You can read '□' as "necessarily" and '◊' as "possibly." But modal logics, like other formal systems, can have many applications. Depending on the application, they might have many different interpretations. For example, we can interpret '$\Box p$' as 'It ought to be the case that p' and '$\Diamond p$' as 'It is permitted to be the case that p'. Table 3.1 lists some different ways in which the modal connectives can be interpreted.

Sometimes the letters 'L' and 'M' are used instead of '□' and '◊'. Sometimes an operator for contingency is also defined:

∇ $\nabla\phi =_{def} (\Diamond\phi \wedge \Diamond\neg\phi)$.

3.1.2 Semantics

Our models for classical propositional logic were just assignments of truth values to the propositional constants—what is represented by the rows of a truth table. Models for modal logic must be more complex.[1] A *model* for modal propositional logic is a quadruple $\langle W, R, @, V \rangle$, where

W is a nonempty set of objects (the *worlds*),

R is a relation defined on W (the *accessibility relation*),

@ is a member of W (the *actual world* of the model), and

[1] The models we describe are called *Kripke models*, after Saul Kripke, who invented them in the late 1950s. There are other ways of doing semantics for modal logic, but this is the most common.

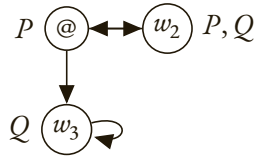


Figure 3.1: A Kripke model. Arrows represent the accessibility relation (here, we have $R@w_2, Rw_2@, R@w_3, Rw_3w_3$). The valuation is represented by writing the atomic formulas that are true at each world next to the world. Thus, in this model $V(P, @)$=$V(P, w_2)$=$V(Q, w_2)$=$V(Q, w_3)$=True, and $V(Q, @)$=$V(P, w_3)$=False.

> V is a function (the *valuation* that assigns a truth value to each pair of a propositional constant and a world.

The triple $\langle W, R, @\rangle$ (without the valuation) is called a *frame*. So you can think of a model as consisting of a frame and a valuation. A model can be represented pictorially, as in Fig. 3.1.

Possible worlds

What are the "worlds"? In many standard applications of modal logic, you can think of the worlds as *possible worlds*—ways things could be, determinate down to the last detail. The valuation function tells us which propositions are true in which possible worlds, that is, which would be true if various counterfactual scenarios obtained. The "actual world" .@ represents the way things *actually* are (according to the model).

There are a lot of controversies about how we should think of possible worlds, metaphysically speaking: whether we should think of them as concrete worlds, as modal realists hold, or as abstract models, properties, or sets of sentences, as modal ersatzists hold.[2] For the most part, we can ignore these controversies when we're just doing logic.

You can think of the accessibility relation as embodying a notion of *relative possibility*. The worlds that are *accessible* from a given world are those that are possible relative to it. If this seems too abstract, you can think of the worlds as seats, and the accessibility relation as the relation that holds between two seats if someone sitting in the first can see the person sitting in the second (Hughes and

[2] If you want to get into this debate, I recommend starting with Stalnaker 1976, then turning to Lewis 1986, David Lewis's magisterial defense of the concretist viewpoint. Loux 1979 collects many relevant papers.

Cresswell 1996). Or think of the worlds as people and the accessibility relation as the relation of loving. When we're considering the logics abstractly, without regard to their applications, it doesn't really matter.

Truth in a model

We define truth in a model for modal formulas in terms of quantification over worlds. Possibility is understood as truth in some accessible world, and necessity as truth in all accessible worlds. Here '$\vDash^{w}_{\mathscr{M}} \phi$' means '$\phi$ is true in model $\mathscr{M}$ at world w'.

- If ϕ is a propositional constant, $\vDash^{w}_{\langle W,R,@,V\rangle} \phi$ iff $V(\phi, w) = \text{True}$.
- $\vDash^{w}_{\langle W,R,@,V\rangle} \ulcorner\neg\phi\urcorner$ iff $\nvDash^{w}_{\langle W,R,@,V\rangle} \phi$.
- $\vDash^{w}_{\langle W,R,@,V\rangle} \ulcorner\phi \wedge \psi\urcorner$ iff $\vDash^{w}_{\langle W,R,@,V\rangle} \phi$ and $\vDash^{w}_{\langle W,R,@,V\rangle} \psi$.
- $\vDash^{w}_{\langle W,R,@,V\rangle} \ulcorner\Diamond\phi\urcorner$ iff for some $w' \in W$ such that Rww', $\vDash^{w'}_{\langle W,R,@,V\rangle} \phi$.
- $\vDash^{w}_{\langle W,R,@,V\rangle} \ulcorner\Box\phi\urcorner$ iff for every $w' \in W$ such that Rww', $\vDash^{w'}_{\langle W,R,@,V\rangle} \phi$.

All the action is in the last two clauses: $\ulcorner\Diamond\phi\urcorner$ says that ϕ is true at some accessible world, while $\ulcorner\Box\phi\urcorner$ says that ϕ is true at every accessible world.

So far we have defined truth in a model *at a world*. We can define (plain) truth in a model in terms of this as follows:

A formula ϕ is true in a model $\langle W, R, @, V\rangle$ if $\vDash^{@}_{\langle W,R,@,V\rangle} \phi$.

This says that a formula is true at a model if it is true at the model's "actual world."

We can now define the logical properties as usual in terms of truth at a model. A sentence is *logically true* if it is true in all models; an argument is *valid* if the conclusion is true in every model in which all the premises are true; and so on.

3.1.3 Modal logics from K to S5

The modal logic K

If we define the logical properties this way and make no further restrictions on what counts as a model, we get the modal logic K. K is the weakest of the modal logics we'll look at, and everything that is valid in K is valid in all the others.

Here are some formulas that are logically true in K:

(2) a. $\Box(P \wedge Q) \supset (\Box P \wedge \Box Q)$ c. $\neg\Box P \equiv \Diamond\neg P$
b. $(\Box P \wedge \Box Q) \supset \Box(P \wedge Q)$ d. $\Box\neg P \equiv \neg\Diamond P$

Can you see why they are true in all models? Think about (2a) and (2b) this way: if '$P \wedge Q$' is true in all accessible worlds, then it must be that 'P' is true in all those worlds and 'Q' is true in all those worlds. The converse also holds: if 'P' is true in all accessible worlds, and so is 'Q', then '$P \wedge Q$' is true in all accessible worlds.

Do you see the resemblance between (2c) and (2d) and the quantifier-negation equivalences? What explains this resemblance?

Here is a formula that is *not* logically true in K:

(3) $\Box P \supset \Diamond P$

Can you see why not? Here is an invalidating model:

Here no world is accessible from @. So '$\Box P$' is trivially true at @, while '$\Diamond P$' is false.

The modal logic D

We get a stronger logic D if we add an additional restriction on models. D-models are K-models that are *serial*.

Serial A relation R on W is *serial* iff $\forall w \in W \exists w' \in W Rww'$.

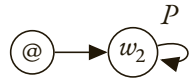
What this means, intuitively, is that there are no "dead ends"—no worlds that can't "see" any worlds (including themselves). Excluding these rules out the countermodel to (3) that we considered above. All instances of '$\Box\phi \supset \Diamond\phi$' are logically true in D.

To say that D is *stronger* than K is to say that every argument valid in K is also valid in D, but there are some arguments that are valid in D but not in K. This goes along with there being some K-models that are not D-models. The more models we admit, the fewer arguments preserve truth in all models.

With dead ends ruled out, (3) no longer has countermodels. However, we can still find a countermodel to

(4) $\Box P \supset P$

Here is one:



What is D good for? Consider the deontic interpretation of the modal operators, where '$\Box$' means *it is obligatory that* and '$\Diamond$' means *it is permissible that*. Thus interpreted, $\ulcorner \Box\phi \supset \Diamond\phi \urcorner$ is the plausible principle that whatever is obligatory is permissible, and $\ulcorner \Box\phi \supset \phi \urcorner$ is the implausible principle that whatever is obligatory is actually the case. Since D validates the former but not the latter, it is a better candidate for the deontic interpretation than K (which invalidates both) or the stronger logics discussed below (which validate both).

The modal logic T

A T-model is a K-model whose accessibility relation is *reflexive*.

Reflexive A relation R on W is *reflexive* iff $\forall w \in W Rww$.

That is, every world can see itself. Since every reflexive accessibility relation is serial, every T-model is a D-model. The converse does not hold: there are D-models that are not T-models. Hence, every argument that is valid in D is valid in T, but not every argument that is valid in T is valid in D.

In T (4) is a logical truth. Indeed, every instance of

(5) $\Box\phi \supset \phi$

is a logical truth. But we can still find a countermodel to

(6) $\Box P \supset \Box\Box P$

(This is left as an exercise.)

The modal logic S4

By imposing another restriction on accessibility relations, we can get a still stronger logic, S4. An S4-model is a K-model whose accessibility relation is both reflexive and *transitive*.

Transitive A relation R on W is *transitive* iff

$$\forall w_0, w_1, w_2 \in W((Rw_0w_1 \wedge Rw_1w_2) \supset Rw_0w_2).$$

Every S4-model is a T-model, so every argument that is valid in T is valid in S4. But some arguments that are valid in S4, such as (6), are not valid in T.

What is S4 good for? Consider the interpretation of '$\Box$' as *it is provable that* and '$\Diamond$' as *it is not refutable that*. Since proofs are mechanically checkable, it is reasonable to think that if something is provable, it is provable that it is provable. (One has only to display the proof and check that it is a proof.) S4 validates all instances of this schema:

(7) $\Box\phi \supset \Box\Box\phi$.

But we might not want the principle that if something is *not* provable, it is provable that it is not provable:

(8) $\neg\Box\phi \supset \Box\neg\Box\phi$.

S4 does not validate this principle.[3]

The modal logic B

We get a different logic, B, if we require that the accessibility relation be both reflexive and *symmetric*.

Symmetric A relation R on W is *symmetric* iff $\forall w, w' \in W (Rww' \equiv Rw'w)$.

B is stronger than T but neither stronger nor weaker than S4. There are arguments valid in B but not S4, and others valid in S4 but not B.

The modal logic S5

Finally, if we require that the accessibility relation be reflexive, symmetric, *and* transitive, we get the modal logic S5. A relation that is reflexive, symmetric, and transitive is called an *equivalence relation*. An equivalence relation partitions the worlds into cells, where each world in a cell can see each other world in that cell (including itself). S5 is strictly stronger than both S4 and B.

It is easy to see that if a formula can be falsified by an S5-model, it can be falsified by a *universal* S5-model—one in which every world is accessible from every other. (Just remove all the cells except the one containing the actual world, and you'll have a universal S5-model that falsifies the same sentences.) So we get the same logic if we think of our models as just sets of worlds, and talk of necessity as truth in all worlds, forgetting about the accessibility relation. You will often

[3] So far so good. However, S4 also validates '$\Box(\Box\bot \supset \bot)$', and hence '$\Box\neg\Box\bot$', which on this reading says that it is provable that a contradiction is not provable. Gödel shows in his second incompleteness theorem that formal systems containing basic arithmetic cannot prove their own consistency in this way. Accordingly, those who work on modal logics of provability generally use logics weaker than S4, which validate (7) but not (5).

find philosophers assuming a picture like this and talking of necessity as truth in *all* possible worlds.

S5 is often thought to be the right logic for logical necessity and metaphysical necessity.

Table 3.2: Summary of the main systems of propositional modal logic.

Logic	Restrictions on accessibility relation	Characteristic axiom
K	—	—
D	serial	$\Box\phi \supset \Diamond\phi$
T	reflexive	$\Box\phi \supset \phi$
S4	reflexive, transitive	$\Box\phi \supset \Box\Box\phi$
B	reflexive, symmetric	$\phi \supset \Box\Diamond\phi$
S5	reflexive, symmetric, transitive	$\Diamond\phi \supset \Box\Diamond\phi$

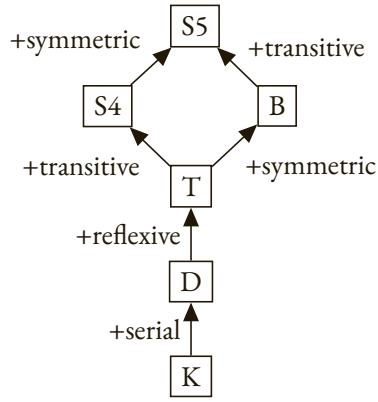


Figure 3.2: Relations between the systems; arrows point to stronger systems.

3.1.4 Proofs

By supplementing our existing proof system for propositional logic with a few rules for the modal operators, we can get new proof systems for T, S4, and S5.

Exercise 3.1: Semantics for modal logics

1. Find a K-model in which '$\Box P \supset \Diamond P$' is false.
2. Find a T-model in which '$\Box P \supset \Box\Box P$' is false.
3. Describe a D-model that is not a T-model.
4. Find an S4-model in which '$\Diamond P \supset \Box\Diamond P$' is false.
5. Find an S4-model in which '$\Diamond\Box P \supset P$' is false.
6. Find a B-model in which '$\Diamond P \supset \Box\Diamond P$' is false.
7. Find an S5-model in which '$\Diamond P \supset \Box P$' is false.
8. Which of these systems would be most appropriate for an interpretation of '$\Box\phi$' as *S knows that* ϕ, and why?

MNE (Modal-Negation Equivalences) You may use the following reversible substitution rules at any point in a proof, citing "MNE" and the line number as justification.

$$\neg\Diamond\phi \iff \Box\neg\phi$$
$$\neg\Box\phi \iff \Diamond\neg\phi$$

Example:

$$\begin{array}{lll} 1 & \neg\Diamond Ga \wedge Gb & \\ \hline 2 & \Box\neg Ga \wedge Gb & \text{MNE 1} \end{array} \tag{3.1}$$

These are modal analogues of the QNE rules from §1.2.

□ Elim If $\ulcorner\Box\phi\urcorner$ occurs in a subproof, you may write down ϕ in the same subproof, with justification '□ Elim'. Example:

$$\begin{array}{lll} 1 & \Box\forall x Fxa & \\ \hline 2 & \forall x Fxa & \Box \text{ Elim } 1 \end{array} \tag{3.2}$$

◊ Intro If ϕ occurs in a subproof, you may write down $\ulcorner \Diamond \phi \urcorner$ in the same subproof, with justification '◊ Intro'. Example:

$$\begin{array}{lll} 1 & \quad Ga \vee Gb & \\ \hline 2 & \quad \Diamond(Ga \vee Gb) & \Diamond \text{ Intro } 1 \end{array} \tag{3.3}$$

Finally, we need some way of *introducing* a '□'.[4] Obviously we can't have the converse of □ Elim, since that would let us prove every instance of $\ulcorner \phi \equiv \Box \phi \urcorner$, and our modal logic would be trivialized.

The trick is to allow '□' to be introduced only through a special kind of subproof, a *modal subproof*, which we will mark with a small box to the left of the subproof line. A modal subproof may be started at any time: there is no separate "Hyp" or "flagging" step. What is special about modal subproofs is that there are strict restrictions on what can be reiterated into them. To reiterate a formula into a modal subproof, we will have to use special, new rules, to be described below. But first, let's see what □ Intro looks like:

□ Intro If your proof contains a modal subproof that ends with a formula ϕ, you can close off the subproof and write $\ulcorner \Box \phi \urcorner$ on the next line, with justification "□ Intro." Schematically:

$$\begin{array}{lll} 1 & \Box & \\ 2 & \quad \vdots & \\ 3 & \quad \phi & \\ 4 & \Box \phi & \Box \text{ Intro } 1\text{–}3 \end{array} \tag{3.4}$$

Note the resemblance to ∀ Intro, which also uses a special subproof. That is hardly an accident, since '□' is semantically a universal quantifier over worlds. In a subproof for ∀ Intro we are reasoning about what is true of an arbitrary object. In a modal subproof we are reasoning about what is true at an arbitrary accessible world.

As we noted above, unrestricted reiteration into modal subproof is not allowed. Thus, we impose the following restriction on our rule Reit:

[4] To avoid complications we will dispense with a ◊ Elim rule, as it isn't necessary given the other rules.

Reit restriction Reit cannot be used to reiterate a formula across a modal subproof boundary (that is, from outside a modal subproof to inside the subproof). Thus, this is okay:

1		P	
2	$\Box$	Q	Hyp
3		R	Hyp
4		Q	Reit 2

(3.5)

But this is not:

1		P		
2	$\Box$	Q	Hyp	
3		P	Reit 1	✗ Illegal!

(3.6)

Here's a way to think about the difference between regular and modal subproofs. Regular subproofs allow things to *enter* freely, but *exit* only according to strict rules. Modal subproofs impose restrictions both on entry *and* on exit. The entry restrictions are given by a *modal reiteration rule*.

Modal Reit T If the subproof immediately containing a modal subproof contains $\ulcorner\Box\phi\urcorner$, you may write down ϕ in the modal subproof.

Modal Reit S4 If the subproof immediately containing a modal subproof contains $\ulcorner\Box\phi\urcorner$, you may write down $\ulcorner\Box\phi\urcorner$ in the modal subproof.

Modal Reit S5 If the subproof immediately containing a modal subproof contains $\ulcorner\Diamond\phi\urcorner$, you may write down $\ulcorner\Diamond\phi\urcorner$ in the modal subproof.

Which modal reiteration rule(s) are available depend on the modal logic. In T, you may use Modal Reit T. In S4, you may use both Modal Reit T and Modal Reit S4. In S5, you may use any of the three rules.

Here's an example of a proof in T of '$(\Box P \wedge \Box Q) \supset \Box(P \wedge Q)$':

1	$\Box P \wedge \Box Q$	
2	$\Box P$	$\wedge$ Elim 1
3	$\Box Q$	$\wedge$ Elim 1
4	$\Box$ P	Modal Reit T 2
5	Q	Modal Reit T 3
6	$P \wedge Q$	$\wedge$ Intro 4, 5
7	$\Box(P \wedge Q)$	$\Box$ Intro 4–6
8	$(\Box P \wedge \Box Q) \supset \Box(P \wedge Q)$	$\supset$ Intro 1–7

(3.7)

Here's a proof in S4 of '$(\Box P \vee \Box Q) \supset \Box(P \vee \Box Q)$':

1	$\Box P \vee \Box Q$	
2	$\Box P$	
3	$\Box$ P	Modal Reit T 2
4	$P \vee \Box Q$	$\vee$ Intro 3
5	$\Box(P \vee \Box Q)$	$\Box$ Intro 3–4
6	$\Box Q$	
7	$\Box$ $\Box Q$	Modal Reit S4, 6
8	$P \vee \Box Q$	$\vee$ Intro 7
9	$\Box(P \vee \Box Q)$	$\Box$ Intro 7–8
10	$\Box(P \vee \Box Q)$	$\vee$ Elim 1, 2–5, 6–9
11	$(\Box P \vee \Box Q) \supset \Box(P \vee \Box Q)$	$\supset$ Intro 1–10

(3.8)

And here's a proof in S5 of '$\Diamond P \supset \Box\Diamond P$':

1	$\Diamond P$	
2	$\Box$ $\Diamond P$	Modal Reit S5 1
3	$\Box\Diamond P$	$\Box$ Intro 2
4	$\Diamond P \supset \Box\Diamond P$	$\supset$ Intro 1–3

(3.9)

Exercise 3.2: Modal natural deductions

1. Use deductions to show that the following arguments are valid in T:

   a) $\dfrac{\begin{array}{l}\Box(P \supset Q)\\ \Box(Q \supset R)\\ \Box(R \supset S)\\ \neg\Diamond S\end{array}}{\neg\Diamond P}$

   b) $\dfrac{\Diamond(P \vee Q)}{\Diamond P \vee \Diamond Q}$

2. Use deductions to show that the following arguments are valid in S4:

   a) $\dfrac{\Diamond\Diamond P}{\Diamond P}$

   b) $\dfrac{\Box(P \wedge \neg Q)}{\Box(\Box P \wedge \neg\Diamond Q)}$

3. Use deductions to show that the following arguments are valid in S5:

   a) $\dfrac{\Box P \vee \Diamond Q}{\Box(P \vee \Diamond Q)}$

   b) $\dfrac{\Diamond\Box P}{\Box P}$

4. For each of the following formulas, determine whether it is a logical truth of T, S4, and/or S5. Give countermodels when a formula is not a logical truth of a system, deductions when it is. (Check each formula against all three systems.)

   a) $\Box(P \supset \Box\Diamond P)$

   b) $\Box\Box P \vee \Box\neg\Box P$

   c) $\Diamond(P \vee Q) \supset \Diamond P$

   d) $\Diamond\Box P \supset \Box P$

   e) $\Diamond\Box\Diamond P \supset \Diamond P$

5. ⋆Thinking about the analogy between □ Intro and ∀ Intro rules, can you come up with an alternative way of stating the ∀ Intro rule that does not require the flagging restriction?

6. ⋆We have given you proof systems for T, S4, and S5. Can you come up with systems that make sense for D and B?

3.2 Modal predicate logic

Recommended reading

W. V. O. Quine, "Reference and Modality" (Quine 1961).

It is natural to try to generalize our modal propositional logic to a modal predicate logic. The interpretation function of a model will need to map predicates to functions from worlds to extensions, since the extension of a predicate can depend on the state of the world. Other decisions are more contentious. Should the interpretation of individual constants be a function from worlds to objects, or just an object? Should the assignment function map variables to objects, or to functions from worlds to objects? And should the domain be relativized to worlds, or constant? One gets different modal predicate logics depending on how these questions are answered.

We won't get as far as answering them in this chapter. Instead, we will focus on some conceptual objections, pressed forcefully by W. V. O. Quine, against the very idea of a modal predicate logic. We must reckon with these before we know how to proceed with the technical development. Readers who want to pursue modal predicate logic in more detail are referred to Hughes and Cresswell 1996.

3.2.1 Opaque contexts

An occurrence of a singular term in a sentence is *purely referential* if all that matters to the truth of the sentence is which object refers to. For example, the occurrence of 'Jack' in

(9) Jack threw a rock

is purely referential. Whether the sentence is true depends only on whether the object denoted by 'Jack' threw a rock. The truth of the sentence does not depend in any way on the *way* we have referred to Jack. Thus, if Jack is Sarah's brother, then (9) is true just in case

(10) Sarah's brother threw a rock

is true. And in general, we can replace the name 'Jack' in (9) with any other singular term that denotes the same object, and we'll get a sentence with the same truth value.

Quine turns this observation into a test for purely referential occurrences. If we can replace an occurrence of a singular term in a sentence with any other singular term that denotes the same object, without affecting the truth value of the containing sentence, then this occurrence of the term is purely referential. On

the other hand, if we cannot always substitute co-referring terms *salva veritate*, then that shows that the occurrence of the term is not purely referential (Quine 1961, p. 140). Quine illustrates his criterion with the sentence

(11) Giorgione was so-called because of his size.

'Giorgione' here denotes the Italian renaissance painter Giorgio Barbarelli da Castelfranco, commonly called Giorgione ("big George"). Although (11) is true, replacing 'Giorgione' with another name for the same painter can yield a falsehood:

(12) Barbarelli was so-called because of his size.

This shows that the occurrence of 'Giorgione' in (11) is not purely referential.

As Quine and other philosophers have noted, there are certain linguistic contexts in which occurrences of terms are generally[5] non-referential. These contexts are called *opaque contexts*. We have already seen one opaque context:

(13) ___ was so-called because of his size.

Other examples include quotational contexts, belief reports, and (at least some) modal contexts:

(14) '___' has six letters.

(15) Lois believes ___ is boring.

Persuade yourself that these contexts are opaque by using Quine's test for purely referential occurrences.

3.2.2 Opaque contexts and quantification

At the core of Quine's objection to quantified modal logic is the observation that, unlike terms like 'Giorgione', variables of quantification can only have purely referential occurrences. Consider

(16) x is so-called because of his size

To make sense of the variable 'x' here, we need to be able to say whether this open formula is *satisfied* by various assignments of objects to 'x'.[6]

Suppose we assign Giorgio Barbarelli da Castelfranco to 'x'. Is the open formula true on that assignment? You might say yes, because it is true that

[5] But not always: see Quine 1961, p. 141.

[6] We are assuming here an objectual interpretation of the variables and quantifiers. As Ruth Barcan Marcus points out (Marcus 1962, p. 258), one way to evade the force of Quine's argument is to interpret the variables and quantifiers substitutionally.

Another way would be to relativize assignments to worlds, or equivalently, assign *individual concepts* (functions from worlds to objects) to the variables. This approach, favored by Carnap, is what Quine has in mind when he argues against the strategy of quantifying over "intensional objects."

Exercise 3.3: Opaque contexts

1. Give an example of an opaque context (besides those mentioned in the text) and show that it is opaque.
2. ⋆We might try to evade the difficulty about 'Giorgione' by thinking of our domain as consisting not of regular objects, but of *pairs* of objects and names? Then we could take (16) to be true of the pair (Giorgione, 'Giorgione') and false of the pair (Giorgione, 'Barbarelli').
   a) If we did this, how could we state the conditions for 'x kicked y' and 'x is so-called' to be satisfied by an assignment of values to the variables?
   b) What could we do with 'There are at least two people who are so-called because of their size?'

(17) Giorgione was so-called because of his size.

But you might equally say no, because it is false that

(18) Barbarelli was so-called because of his size.

So there is no good way to answer the question whether the open sentence (16) is true on an assignment that assigns the painter to x.

This shows, Quine thinks, that it doesn't make sense to quantify into opaque contexts like the one generated by 'so-called'. It makes sense for names (like 'Giorgione') to have occurrences that aren't purely referential, but variables can only be purely referential and can't occur in opaque contexts.

3.2.3 The number of planets argument

To argue against the possibility of quantified modal logic, then, Quine just has to show that modal operators create opaque contexts. Since (he has argued) variables do not make sense in opaque contexts, formulas like

(19) $\exists x \Box Fx$.

are not intelligible.[7]

[7] Of course, one could still allow modal operators to occur in front of formulas without free variables. But the resulting logic would not allow us to express anything we could not express using a necessity predicate (Quine 1976).

To show that '□ ___ ' is an opaque context, it suffices to show that substitution of co-denoting terms in the position occupied by ' ___ ' does not preserve truth value. Quine notes first that

(20) $\Box(9 > 5)$

is a true sentence. It is a matter of mathematical necessity that 9 is greater than 5. However, '9' denotes the same number as 'the number of planets'.[8] And

(21) $\Box$(the number of planets > 5)

is false. That there are more than five planets is a matter of contingent fact. Had conditions in the early solar system been different, we might have had fewer than five planets. This shows, Quine argues, that the '□' operator creates an opaque context. Since variables do not make sense inside opaque contexts, '□' cannot sensibly prefix an open formula. Thus formulas like

(22) $\Box(x > 5)$

or of

(23) $\exists x \Box(x > 5)$.

are unintelligible.

3.2.4 Smullyan's reply

Smullyan (1948) thinks that Quine's argument can be met. Smullyan assumes, with Russell, that definite descriptions can be understood as quantifiers. But that means they have scopes. He thinks that Quine's argument ignores this, and trades on a scope ambiguity.

Using modern generalized quantifier notation, we can put Smullyan's point like this. Quine's sentence (21) is ambiguous between

(24) $the_x(Nx, \Box(x > 5))$

and

(25) $\Box the_x(Nx, x > 5)$

(where Nx = 'x numbers the planets'). In (24), the necessity operator takes narrow scope with respect to the description; in (25), it takes wide scope. (24) does follow logically from (20) and

[8] We now take there to be eight planets, but at the time Quine was writing, Pluto was considered a planet.

Exercise 3.4: Quine and Smullyan

Give a deduction of (24) from premises (26) and (20). You may use all of the rules for propositional modal logic and standard predicate logic. Try to see why you can't get (25) from these premises.

(26) $the_x(Nx, x{=}9)$
The number of planets is 9,

while (25) does not.

Smullyan takes these considerations to show that, in applying Frege's "substitutivity" criterion for referential occurrences of terms, we should restrict ourselves to genuine *names*. If we treat definite descriptions like names, ignoring their scopes, then we will run into modal paradoxes. But these paradoxes "arise not out of any intrinsic absurdity in the use of modal operators but rather out of the assumption that descriptive phrases are names" (Smullyan 1948, p. 34).

Quine acknowledges the point (Quine 1976, p. 173; Quine 1961, p. 154). He doesn't think it would help to use examples with proper names, because on his view names can be definitionally reduced to definite descriptions (Pegasus = the thing that pegasizes). But he thinks we can see the problem even in pure quantification theory, without names or descriptions. It is a theorem of first-order logic with identity that

(27) $\forall x \forall y(x{=}y \supset \phi x \equiv \phi y)$

for any open formula ϕ. So, let 'ϕ ___' be '$\Box x{=}$ ___'. Then

(28) $\forall x \forall y(x{=}y \supset \Box x{=}x \equiv \Box x{=}y)$

But

(29) $\forall x \Box x{=}x$

is a theorem, so we can derive

The necessity of identity $\forall x \forall y(x{=}y \supset \Box x{=}y)$

The upshot is that, if it makes sense to quantify inside modal contexts, we must accept that if $x{=}y$, then $\Box x{=}y$.

Quine thinks this is absurd, since there are clearly some true identities that are true only contingently. For example, Hesperus is identical with Phosphorus,[9]

[9] Both 'Hesperus' (Evening Star) and 'Phosphorus' (Morning Star) are names for the planet Venus.

but, Quine thinks, this is a matter of contingent fact—something that the ancient astronomers had to discover through painstaking observation. More generally, Quine rejects the idea that there are ϕ and ψ such that:

(30) $\exists x(\phi x \wedge \psi x \wedge \Box\phi x \wedge \neg\Box\psi x)$

To accept an instance of this schema is to enter the "metaphysical jungle of Aristotelian essentialism" (Quine 1976, p. 176):

> This is the doctrine that some of the attributes of a thing (quite independently of the language in which the thing is referred to, if at all) may be essential to the thing, and others accidental (Quine 1961, p. 155).

According to Aristotelian essentialism, an object—Fido, for example—has certain properties necessarily (being a dog, being an animal), and others only contingently (being well-trained, weighing 50 pounds). On Quine's view, this is dark metaphysics that be made intelligible in the modern scientific worldview. We will soon see Kripke defending quantified modal logic by giving a philosophical defense of both the necessity of identity and Aristotelian essentialism.

3.3 The slingshot argument

Quine sometimes appeals to a different technical argument (Quine 1961, p. 159; see also Quine 1976, p. 163). This argument has come to be called "the slingshot," in recognition of its small size and giant-slaying potential.[10] The argument purports to show that any context that admits both substitution of identicals and substitution of logical equivalents is truth-functional. Quine takes this to be an impossibility proof for quantified modal logic. He has already argued that variables are not intelligible inside an opaque context (§3.2.2). He takes the slingshot to show that all non-truth-functional operators create opaque contexts. If this is right, then no non-truth-functional operator can intelligibly be applied to formulas containing variables.

We assume that the following rules are valid for the $\Box$ operator:

Substitution of logical equivalents (Equiv) If ϕ and ψ are logically equivalent, then $\ulcorner\Box\psi\urcorner$ may be inferred from $\ulcorner\Box\phi\urcorner$.

Substitution of co-denoting definite descriptions (Coden) From $\ulcorner t_1{=}t_2\urcorner$ and $\ulcorner\Box\phi\urcorner$, where t_1 and t_2 are definite descriptions, $\ulcorner\Box\phi^{t_2/t_1}\urcorner$ may be inferred, where ϕ^{t_2/t_1} is the result of substituting t_2 for t_1 in ϕ.

[10] The basic idea of a slingshot argument was first proposed by Alonzo Church (Church 1943). The presentation here, which differs from the version in Quine 1961 in using definite descriptions rather than class abstracts, is indebted to Neale 1995; Neale 2001.

We then argue as follows for any sentences ϕ and ψ:

1	$\phi \wedge \psi$	
2	$\quad \Box\phi$	
3	$\quad \Box(a{=}\imath x(x{=}a \wedge \phi))$	Equiv 2
4	$\quad \imath x(x{=}a \wedge \phi){=}\imath x(x{=}a \wedge \psi)$	(provable from 1)
5	$\quad \Box(a{=}\imath x(x{=}a \wedge \psi))$	Coden 3, 4
6	$\quad \Box\psi$	Equiv 5
7	$\quad \Box\psi$	
8	$\quad \vdots$	(as above)
9	$\quad \Box\phi$	
10	$\Box\phi \equiv \Box\psi$	$\equiv$ Intro 2–9

This argument establishes the conditional

$$(\phi \wedge \psi) \supset (\Box\phi \equiv \Box\psi).$$

That is: if ϕ and ψ are both true, then $\ulcorner\Box\phi\urcorner$ and $\ulcorner\Box\psi\urcorner$ have the same truth value. To show that '$\Box$' is truth-functional, we must also show that if ϕ and ψ are both false, then $\ulcorner\Box\phi\urcorner$ and $\ulcorner\Box\psi\urcorner$ have the same truth value. We can do that by means of the following argument (Correia 2003, p. 441):

1	$\neg\phi \wedge \neg\psi$	
2	$\quad \Box\phi$	
3	$\quad \Box\neg(a{=}\imath x(x{=}a \wedge \neg\phi))$	Equiv 2
4	$\quad \imath x(x{=}a \wedge \neg\phi){=}\imath x(x{=}a \wedge \neg\psi)$	(provable from 1)
5	$\quad \Box\neg(a{=}\imath x(x{=}a \wedge \neg\psi))$	Coden 3, 4
6	$\quad \Box\psi$	Equiv 5
7	$\quad \Box\psi$	
8	$\quad \vdots$	(as above)
9	$\quad \Box\phi$	
10	$\Box\phi \equiv \Box\psi$	$\equiv$ Intro 2–9

Together these two arguments suffice to show that

'$\Box$' is truth-functional $(\phi \equiv \psi) \supset (\Box\phi \equiv \Box\psi)$.

3.3.1 Applications of slingshot arguments

We have presented the slingshot argument with the modal box '$\Box$', and thus as an argument against quantified modal logic. However, we can also consider different interpretations of '$\Box$', yielding different applications of the slingshot.

Facts Read $\ulcorner\Box\phi\urcorner$ as $\ulcorner$the fact that ϕ = the fact that snow is white$\urcorner$. Then the conclusion is that every fact is identical to the fact that snow is white—that is, there is only one fact ("The Great Fact," as Donald Davidson calls it). Davidson takes this to be a refutation of the view that for a statement to be true is for it to correspond to a fact (Davidson 1984, pp. 41–2).

Propositions Read $\ulcorner\Box\phi\urcorner$ as $\ulcorner$the proposition that ϕ = the proposition that $1 > 0\urcorner$. Then the conclusion is that there is just one true proposition.

Belief Read $\ulcorner\Box\phi\urcorner$ as $\ulcorner$Albert believes that $\phi\urcorner$. Then the conclusion is that if Albert believes one true proposition, he believes every true proposition (Quine 1960, pp. 148–9).

In each case, the slingshot argument forces us to give up either substitution of logical equivalents or substitution of co-denoting definite descriptions in the contexts in question.

3.3.2 The Gödel slingshot

Gödel gave a different version of the argument (Gödel 1944) which replaces *Equiv* with a weaker principle,

Gödel equivalence (Gödel) From $\ulcorner\Box\Phi\alpha\urcorner$, $\ulcorner\Box(\alpha{=}\imath x(x{=}\alpha \wedge \Phi x))\urcorner$ may be inferred, and vice versa (where Φ is a predicate and α a term).

The argument runs as follows:

1	$a{\neq}b \wedge Fa \wedge Gb$	
2	$\Box Fa$	
3	$\Box(a{=}\imath x(x{=}a \wedge Fx))$	Gödel 2
4	$\imath x(x{=}a \wedge x{\neq}b){=}\imath x(x{=}a \wedge Fx)$	(provable from 1)
5	$\Box(a{=}\imath x(x{=}a \wedge x{\neq}b))$	Coden 3,4
6	$\Box a{\neq}b$	Gödel 5
7	$\Box(b{=}\imath x(x{=}b \wedge a{\neq}x))$	Gödel 6
8	$\imath x(x{=}b \wedge a{\neq}x){=}\imath x(x{=}b \wedge Gb)$	(provable from 1)
9	$\Box(b{=}\imath x(x{=}b \wedge Gb))$	Coden 7, 8
10	$\Box Gb$	Gödel 9
11	$\Box Gb$	
12	⋮	(as above)
13	$\Box Fa$	
14	$\Box Fa \equiv \Box Gb$	≡ Intro 2–13

The difference between the Quine and Gödel slingshots is mainly interesting in connection with theories of facts and propositions. As Barwise and Perry (1981) pointed out, logical equivalents can bring in new material that you might think becomes "part of" the proposition. For example,

(31) John is sleeping and either Bert is awake or Bert is not awake

seems to be about Bert in a way that

(32) John is sleeping

is not. The Gödel slingshot might be preferable, then, for arguing against theories of facts and propositions, because it does not require *Equiv*.

3.3.3 Critique of the slingshot

To evaluate the slingshot argument, we need to assess its premises, *Coden* and *Equiv* (or, in Gödel's version, *Coden* and *Gödel*). To do that, we must interpret '□' in a particular way, and ask whether the principles hold on that interpretation.

Exercise 3.5: The slingshot argument

1. Prove that 'P' is logically equivalent to '$a{=}\imath x(x{=}a \wedge P)$'. Use the Russellian Equivalences from §2.2.4. (To prove logical equivalence, it suffices to prove a biconditional using no premises.)
2. Show that '$\imath x(x{=}a \wedge P){=}\imath x(x{=}a \wedge Q)$' can be derived from '$P \wedge Q$' using standard logical rules and the Russellian Equivalences. (Note that the definite descriptions take narrow scope.)
3. ⋆Suppose we wanted to say (against Russell) that when there is not a unique x such that Fx, '$G(\imath xFx)$' is neither true nor false. Would 'P' still be logically equivalent to '$a{=}\imath x(x{=}a \wedge P)$'? Why or why not? (Careful: In this case the Russellian Equivalences could not be used. You might try thinking of rules for $\imath$ that would still be valid. What would validity mean in such a system?)

Equiv is certainly plausible when '$\Box$' is interpreted as a necessity operator. If ϕ is necessary and ψ is logically equivalent to ϕ, then surely ψ is also necessary. As we have seen, *Equiv* becomes less compelling when we apply the slingshot to theories of facts or propositions, but *Gödel* might then provide a plausible replacement.

Coden is more contentious. Recall that there are two schools of thought about definite descriptions: (1) they are singular terms, like names; (2) they are quantifiers. If definite descriptions are quantifiers, as Russell held, then it is easy to see why ϕ is logically equivalent to $\ulcorner a{=}\imath x(x{=}a \wedge \phi)\urcorner$. But then co-denoting definite descriptions cannot be intersubstituted, as Smullyan reminds us, unless the definite description takes wide scope. And in the slingshot, the modal operator has wide scope over the definite description.

Does the argument fare better if definite descriptions are singular terms, as Frege held? In that case, perhaps, *Coden* is on firmer ground. But then it is harder to see why our statements should be logically equivalent. (Exercise 3.5 (3) explores this issue.)

If the slingshot argument establishes anything beyond a doubt, it is the importance of studying philosophical logic. Some of the most influential philosophers of the twentieth century used the argument to justify controversial philosophical and methodological stances. So it is a matter of some importance whether the argument is sound. But determining that turns on delicate issues about scope, quantification, and modality.

3.4 Kripke's defense of *de re* modality

Recommended reading

Saul Kripke, *Naming and Necessity* (Kripke 1980, pp. 34–63, 97–105).

3.4.1 Kripke's strategy

We saw Quine argue that quantified modal logic is only intelligible if "Aristotelian essentialism" is true—that is, if objects have some properties necessarily, others contingently, independent of how they are described. Quine thought we might be able to make sense of necessity *de dicto*—as a property of what is said, and hence dependent on how we describe things. The *dictum* 'the second planet from the sun, if there is one, is a planet' is necessary because, if our subject can be picked out as 'the second planet from the sun', it must be a planet. Just by thinking about what it takes for this *dictum* to be true, we can see that it must be true no matter how the world is arranged. For similar reasons it is necessary that the number that numbers the planets numbers the planets. But Quine thought it was dark metaphysical nonsense to suppose that there could be modality *de re*—as a property of things, independently of how they are described:

> ...necessity does not properly apply to the fulfillment of conditions by *objects* (such as the ball of rock which is Venus, or the number which numbers the planets), apart from special ways of specifying them. (Quine 1961, p. 151)

If we consider Venus (or the number 8) independently of any particular way of describing it, Quine thinks, it makes no sense to ask what properties it has necessarily. Necessity, for Quine, is a creature of language, a reflection of the way we describe things.

Saul Kripke's strategy for vindicating quantified modal logic is to argue that we have a perfectly ordinary, intuitive grasp of *de re* modality. It's not a pernicious idea introduced by metaphysicians. To defend it, we just need to dispel some philosophical confusions that have made it *seem* incoherent:

> I don't know if some philosophers have not realized this; but at any rate it is very far from being true that this idea [that a property can meaningfully be held to be essential or accidental to an object independently of its description] is a notion which has no intuitive content, which means nothing to the ordinary man. Suppose that someone said, pointing to Nixon, 'That's the guy who might have lost'. Someone else says 'Oh no, if you describe him as "Nixon", then he might have lost; but, of course, describing him as the winner, then it is not true that he might have lost'. Now which one is being the philosopher, here, the unintuitive man? It seems to me obviously to be the second. (Kripke 1980, p. 41)

Why have philosophers resisted the idea, which Kripke finds so natural, that objects can have modal properties independently of how they are described? The root reason, he thinks, is that they have assumed that *necessity* and *a priority* go together.[11] This assumption is accepted by Quine, who shares the logical empiricists' view that all necessity must be rooted in language and logic. But it is also accepted by many other philosophers: for example, Kant, who rejects the idea that all necessity is due to logic and definitions, explicitly affirms that necessity and a priority go together (Kant 1965, B4).[12]

If we accept that whatever is necessary is knowable a priori, and whatever is knowable a priori is necessary, it becomes very hard to make sense of *de re* modality. For we then have to accept that

(33) Tully = Cicero

is not necessary because it's not knowable a priori. On the other hand,

(34) Cicero = Cicero

is necessary because it's knowable a priori. So, how can we make sense of

(35) Necessarily, x = Cicero?

What is its truth value when we assign the man known as both 'Cicero' and 'Tully' to x? One wants to say: it is senseless to ask whether this man is necessarily identical to Cicero. Described as 'Cicero', he is; described as 'Tully', he is not.

To remove this obstacle, Kripke is going to argue that some contingent truths can be known a priori, and that some necessary truths are knowable only a posteriori (on the basis of experience). This clarification will remove the philosophical impediments to understanding *de re* modality.

3.4.2 The contingent a priori

Ludwig Wittgenstein says this about the meter bar:[13]

[11] To say that a proposition is knowable a priori is to say that one can know it on grounds that are independent of sensory experience. Traditionally, the truths of mathematics and logic were assumed to be paradigms of truths knowable a priori. By contrast, the proposition that there are planets larger than the earth is knowable only a posteriori, on the basis of experiences such as telescopic observations.

[12] The ancient philosophical tradition is another matter. Aristotle holds that the first principles of *all* sciences—even, for example, biology—must be necessary (*Posterior Analytics* I.1).

[13] This was back when the meter standard was definitive. Today, a meter is defined as the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second. A second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom. So in effect, we have substituted the cesium-133 atom for the standard meter bar. The same points could still be made, but we'll stick with the meter bar for simplicity.

> There is one thing of which one can say neither that it is one meter long nor that it is not one meter long, and that is the standard meter in Paris. But this is, of course, not to ascribe any extraordinary property to it, but only to mark its peculiar role in the language game of measuring with a meter rule. (Wittgenstein 1958, §50)

Kripke (1980, p. 54) objects: "If the stick is a stick, for example, 39.37 inches long (I assume we have some different standard for inches), why isn't it one meter long?"

The interesting question is, why does Wittgenstein say this odd thing? Well, if the standard meter is one meter long, then it seems we can know that it is one meter long without measuring it. Our stipulation makes it the case that it's one meter long. Since this stipulation is what fixes the meaning of 'meter', we can know a priori, purely by reflecting on the rules for use of the word 'meter', that the standard meter is one meter long.

That's odd, given the traditional view that what is a priori is necessary, because it doesn't seem that the meter is *necessarily* one meter long. It *could have been* shorter, or longer, than it is, and in that case it would have been shorter, or longer, than a meter.

You might think: ah yes, but then a meter would have also been shorter, or longer, so the standard meter bar would still have been a meter long. Kripke dismisses this line of thought as flawed. If the meter bar had been shorter, we would have *called* shorter things "one meter long," but that is not to say that they would have *been* one meter long. What we're imagining is a scenario where we used the word 'meter' to denote lengths of less than a meter.[14]

It's important here that 'one meter' is not introduced as a *synonym* for 'the length of the standard meter bar'. If it were, it would pick out different lengths in different possible worlds. Rather, it is introduced as the name for a *particular length*, the same in all possible worlds. We identify this length by pointing to the standard meter bar. The meter bar serves to fix the *reference* of 'meter', not to give its *meaning*. (Suppose we discover a new island, and say: "we hereby dub the island we're standing on 'Newlandia'." When we move on and discover yet another new island, 'Newlandia' is still the name of the island we were originally standing on. 'Newlandia' does not mean *the island we are standing on*.)

[14] The following riddle is often attributed to Abraham Lincoln: "You remember the slave who asked his master, 'If I should call a sheep's tail a leg, how many legs would it have?' 'Five.' 'No, only four; for my calling the tail a leg would not make it so.' " Lincoln was using the riddle to resist proclaiming the slaves free during the civil war: "if I say to the slaves, 'you are free,' they will be no more free than at present." (*Daily Milwaukee News*, September 23, 1862, p. 1 col. 2. Quoted at `http://quoteinvestigator.com/2015/11/15/legs/` which traces the riddle back to 1825.)

So, Kripke concludes, it seems to be knowable a priori that the standard meter (if it exists) is 1 meter long, even though this is a contingent fact:

> What then, is the *epistemological* status of the statement 'Stick S is one meter long at t_0', for someone who has fixed the metric system by reference to Stick S? It would seem that he knows it *a priori*. For if he used stick S to fix the reference of the term 'one meter', then as a result of this kind of 'definition' (which is not an abbreviative or synonymous definition), he knows automatically, without further investigation, that S is one meter long. On the other hand, even if S is used as the standard of a meter, the *metaphysical* status of 'S is one meter long' will be that of a contingent statement, provided that 'one meter' is regarded as a rigid designator: under appropriate stresses and strains, heatings or coolings, S would have had a length greater than one meter even at t_0. (Kripke 1980, p. 56)

By *rigid designator*, Kripke just means a term that denotes the same thing (here, the same length) with respect to every possible situation.

For similar reasons, I can know a priori that I am here, and that I am *this* tall (putting my hand on my head). Both of these propositions are contingent: I could have been elsewhere, and I could have been taller. But I do not need to do any empirical observation to know these things.

3.4.3 The necessary a posteriori

Conversely, Kripke argues, there can be necessary truths that are knowable only a posteriori. An example is 'Hesperus = Phosphorus':

> Are there really circumstances under which Hesperus wouldn't have been Phosphorus? Supposing that Hesperus is Phosphorus, let's try to describe a possible situation in which it would not have been. Well, it's easy. Someone goes by and he calls two *different* stars 'Hesperus' and 'Phosphorus'. It may even be under the same conditions as prevailed when we introduced the names 'Hesperus' and 'Phosphorus'. But are those circumstances in which Hesperus is not Phosphorus or would not have been Phosphorus? It seems to me that they are not. (Kripke 1980, p. 102)

The scenario Kripke has imagined is not one in which Hesperus isn't Phosphorus; it's one in which the *names* 'Hesperus' and 'Phosphorus' don't denote the same things they do in the actual world.

To persuade us of this, Kripke argues 'Hesperus' and 'Phosphorus', and proper names generally, are rigid designators. Their denotation does not vary across possible worlds. It follows from this that identities between rigid designators will be necessary.

Why think names are rigid designators? Kripke offers an "intuitive test" (Kripke 1980, p. 48): Ask whether someone other than X might have been X. If the

answer is no, then *X* is a rigid designator. If the answer is yes, then *X* is not a rigid designator. Let's apply this test to a definite description and a name, respectively:

(36) Someone other than *the US President in 1970* might have been *the US President in 1970*. [YES ⇒ non-rigid]

(37) Someone other than *Nixon* might have been *Nixon*. [NO ⇒ rigid]

> ...although the man (Nixon) might not have been the President, it is not the case that he might not have been Nixon (though he might not have been *called* 'Nixon'). (Kripke 1980, p. 49)

In this way, Kripke defends what Quine took to be a *reductio ad absurdum* of quantified modal logic: the necessity of identity (30).

At the end of §3.4.1, we presented the following argument in the spirit of Quine:

1. 'Tully = Cicero' is not knowable a priori.
2. 'Cicero = Cicero' is knowable a priori.
3. Something is necessary iff it is knowable a priori.
4. By 1 and 3, 'Tully = Cicero' is not necessary.
5. By 2 and 3, 'Cicero = Cicero' is necessary.
6. 'Tully' and 'Cicero' denote the same object.
7. By 4, 5, and 6, there is no way to make sense of the open formula 'necessarily x = Cicero'.

We can now see that this argument can be resisted if, following Kripke, we give up premise 3. It is then open to us to hold that 'Tully = Cicero' and 'Cicero = Cicero' are both necessary truths, and that the open sentence 'necessarily x = Cicero' holds of the object Cicero (a.k.a. Tully), no matter how we denote it.

3.4.4 Epistemic and alethic modals

In ordinary language, 'must' often has an epistemic sense, where it is roughly equivalent to *given what I know, it follows that*:[15] Consider

(38) Smith must be the murderer.

Since what I know could imply that Smith is the murderer without implying that Dr. Zero is the murderer, even if Smith *is* Dr. Zero, (38) can be true even when

[15] For more on this "roughly," see MacFarlane 2014, ch. 10.

(39) Dr. Zero must be the murderer.

is false. For this kind of modal operator, the necessity of identity is implausible, and quantifying in will be problematic.

Kripke acknowledges that modal words are sometimes used in this epistemic way. Suppose we don't know yet that Hesperus is Phosphorus. We might say: 'it's possible that Hesperus is Phosphorus, and it's possible that it isn't', or 'Hesperus might be Phosphorus, and it might not be.' We're using 'possible' and 'might' here in the epistemic sense.

What Kripke wants to show us is that we *do* understand, and constantly use, a sense of modal words that isn't epistemic. (Rather unfortunately given his aims, he calls this sense *metaphysical*; the term *alethic* is also used.) And for *this* sense of the modal terms, it makes sense to quantify in.

However, for non-alethic interpretations of the modal operators, Quine's argument may still be a good one. If we interpret the box as 'it is logically necessary that', then it doesn't make much sense to quantify inside the box. For '$a=a$' is a logical truth, but '$a=b$' isn't, even when 'a' and 'b' denote the same object. So there's no making sense of '$a=x$'. A similar conclusion will hold for any sense of necessity that is connected with apriority, like 'It is analytic that...' or 'It is knowable a priori that...'. Quine assumed that the modal '□' operator would have to be read in one of these ways, and on that assumption his criticism is valid. Kripke's point is that there's something else we can mean by '□', something that supports quantifying in, and that it's not a dark metaphysical fantasy, but something very ordinary. Philosophy is not needed to give us an understanding of alethic modality, but only to remove the philosophical impediments to recognizing that we understood it all along.

Further readings

- Hughes and Cresswell 1996 is a good technical introduction to modal logic. Girle 2000 is useful for its chapters on natural deductions and philosophical applications.
- See Quine 1976 for further elaboration of Quine's objections to the intelligibility of quantified modal logic. For some alternative ways of making sense of quantified modal logic, see Carnap 1956 (§44) and Lewis 1968.
- Smullyan 1948 sets out Smullyan's objection to Quine. (Note that this article uses the old-fashioned notation of Russell and Whitehead's *Principia Mathematica*. In this notation, dots sometimes mean conjunction and are sometimes used instead of parentheses to indicate grouping, and subscripted

variables express quantification. For an explanation of how to read this notation, see Linsky 2016.)

- On the slingshot argument, see Neale 1995.

4 Conditionals

4.1 The material conditional

Recommended reading

James F. Thomson, "In Defense of '⊃' " (Thomson 1990). (You may skip the discussion of Strawson on pp. 61–64.)

In introductory logic, you were taught to formalize English conditionals using a truth-functional connective, the material conditional ('⊃'). You were probably also taught that this wouldn't always give good results, and that it should be considered a simplification that is useful for some purposes.

In this chapter we will ask three interrelated questions:

- Can we come up with better truth conditions for conditionals?
- If not, why not? Is it because conditionals are truth-functional after all, or because they don't have truth conditions at all?
- What logical principles hold for conditionals?

4.1.1 Indicative vs. counterfactual

It is commonly accepted that there are two fundamentally different varieties of conditionals in English (and other natural languages), *indicative* and *subjunctive*.

In English and many other languages, the distinction is marked by differences in grammatical mood. In indicative conditionals, the verb in the consequent is in the indicative mood, while in subjunctive conditions, it is in the subjunctive mood. Here is a nice minimal pair that shows the difference:

(1) If Oswald didn't shoot Kennedy, someone else did. [indicative]

(2) If Oswald hadn't shot Kennedy, someone else would have. [subjunctive]

To see the difference, ask yourself what would be evidence for each. If we think that Kennedy was, in fact, shot, then we'll accept (1) without any additional

evidence. We know *someone* shot him. If it wasn't Oswald, then it must have been someone else.

By contrast, knowing that Kennedy was shot is not sufficient for accepting (2). You might accept (2) if you had evidence that Oswald didn't act alone but was part of a larger conspiracy, or if you think that so many people wanted to kill Kennedy that someone else would have stepped up. But if you think Oswald was acting alone, and an anomaly, you'd reject (2).

Because it could be rational to accept (1) and reject (2), they would seem to have different truth conditions. (2) concerns what would have happened in a possible scenario where Oswald didn't shoot Kennedy. (If in fact Oswald did shoot Kennedy, then this scenario is a counterfactual one, an alternative "possible world.") (1), by contrast, concerns what really happened in the world as it is.

You'll often hear subjunctive conditionals called *counterfactual conditionals*, because their antecedents are being presented as contrary to actual fact. But don't think that a counterfactual conditional is a conditional with a false antecedent, and an indicative a conditional with a true antecedent. A counterfactual conditional can have a true antecedent. (Suppose you mistakenly think you forget to turn in the last homework assignment, and you say: 'If I had turned in the last homework assignment, I would have passed the class'.) And an indicative conditional can a false antecedent, as in (1) above. That said, it would be weird to assert a counterfactual that you *knew* had a true antecedent, or an indicative that you *knew* had a false antecedent, and any account of the difference between indicatives and subjunctives should offer some explanation of this fact.

It seems pretty clear that subjunctive conditionals aren't material conditionals. After all, we generally use them when we know the antecedent is false. In those cases the corresponding material conditional is *always* true, but we have pretty clear intuitions that some of the subjunctive conditionals are false. Consider these pairs:

(3) a. If I were seven feet tall, I could change the light bulb. (T)
    b. If I were four feet tall, I could change the light bulb. (F)

(4) a. If I had dropped this pencil, it would have stayed on the ground. (T)
    b. If I had dropped this pencil, it would have bounced to the ceiling. (F)

When it comes to indicative conditionals, the material conditional analysis cannot so easily be dismissed. For, as noted above, we simply don't use material conditionals with antecedents that we take to be false. If I were to say

(5) If I am four feet tall, I can change the light bulb,

you would be puzzled about what I am trying to express, rather than having a clear judgment of falsity. (Am I ignorant of my own height?)

Take out a piece of paper. For each of the following sentences, write 'T' if you think it is true, 'F' if you think it is false. If you think it's wrong to call a particular sentence either true or false, you can put '?'.

(6) a. If snow is black, North Korea is in Europe.
b. If snow is black, North Korea is in Asia.
c. If snow is white, North Korea is in Asia.
d. If snow is white, North Korea is in Europe.

Think about the pattern of answers you gave, and ask others how they answered these questions. What, if anything, do these answers show?

4.1.2 Entailments between indicatives and material conditionals

One way to get clearer about the truth conditions of indicatives is to ask about entailments. In what follows, we will use the symbol '$\supset$' for the material conditional, and '$\rightarrow$' for the indicative conditional.

Nearly everyone accepts that the indicative conditional entails the material conditional:

(7) $$\frac{p \rightarrow q}{p \supset q}$$

This inference must be valid if Modus Ponens is valid for the indicative conditional. For suppose $\ulcorner p \rightarrow q \urcorner$ were true and $\ulcorner p \supset q \urcorner$ false. Then we'd have a counterexample to Modus Ponens for the indicative conditional, for the falsity of $\ulcorner p \supset q \urcorner$ requires that p be true and q false. In §4.4 we'll look at an argument that Modus Ponens in fact *fails* for the indicative conditional. But if we want our conditional to respect Modus Ponens, we'd better accept the inference (7).

More controversial is the converse entailment:

(8) $$\frac{p \supset q}{p \rightarrow q}$$

If both this and (7) were valid, we could show that $\ulcorner p \rightarrow q \urcorner$ is *equivalent* to $\ulcorner p \supset q \urcorner$, and the material conditional analysis of indicatives would be vindicated. Those who think that the indicative conditional is not a material conditional therefore reject (8). They agree that in order for $\ulcorner p \rightarrow q \urcorner$ to be true, $\ulcorner p \supset q \urcorner$ must be true, but they think that some additional connection between antecedent and consequent is required as well.

This "received opinion" is the target of Thomson 1990.

4.1.3 Thomson against the "received opinion"

Thomson makes three important observations:

- If the received opinion is correct, conditionals with false antecedents or true consequents that lack the requisite connection between antecedent and consequent are just *false*. But although we're reluctant to call them true, calling them false doesn't seem right either (Thomson 1990, p.59).
- Even if assertions of $\ulcorner p \rightarrow q \urcorner$ typically *communicate* that there is some non-truth-functional connection between p and q, it does not follow that such a connection is required for the *truth* of $\ulcorner p \rightarrow q \urcorner$.
- It may be that it is bad *reasoning* to move from $\ulcorner \neg p \urcorner$ to $\ulcorner p \rightarrow q \urcorner$. But that does not mean that $\ulcorner \neg p \urcorner$ does not *entail* $\ulcorner p \rightarrow q \urcorner$.

These last two points stand in need of more explanation.

What is said vs. what is implied

If I assert a disjunction, my audience will generally assume that I don't know which disjunct is true, since if I did, I would have made the stronger assertion. Otherwise I'd be uncooperative, and it's generally assumed that conversational partners won't withhold relevant information. For example, if I say

(9) Sam is either at the bar or studying.

you'll assume I don't know that Sam is at the bar. (This point is due to Grice 1989.)

If indicative conditionals are material conditionals, they are equivalent to disjunctions: $\ulcorner p \supset q \urcorner$ is logically equivalent to $\ulcorner \neg p \vee q \urcorner$. So

(10) If Sam is not at the bar, he is studying,

is logically equivalent to (9), and the point we just made about (9) applies here too. If I assert (10), then the normal implication is that I don't know the antecedent or the consequent. (For if I did, I'd have been in a position to make a more informative assertion, and as a cooperative conversation partner, I would have done so.) If I have reason for thinking that (10) is true, but I don't know the truth value of the antecedent or the consequent, that must be because I know something about the *relation* of the antecedent to the consequent. So, when I assert (10), others will reasonably assume that I take there to be some relation between whether Sam is at the bar and whether he is studying.

In this way, Thomson thinks, we can explain why assertions of conditionals typically communicate that there is some non-truth-functional relation between

antecedent and consequent, even though the truth of the conditional doesn't require any such relation: "we read what we take to be [the speaker's] reason into the statement itself" (Thomson 1990, p. 68).

Good reasoning vs. entailment

We can all agree that inferring $\ulcorner p \rightarrow q \urcorner$ from $\ulcorner \neg p \urcorner$ looks like bad reasoning. But does it follow that $\ulcorner \neg p \urcorner$ does not *entail* $\ulcorner p \rightarrow q \urcorner$? Thomson argues that it does not. Reasoning from a premise to one of its logical consequences can sometimes be bad reasoning.

In the example at hand, this is because the conclusion is junk. If one is asserting $\ulcorner p \rightarrow q \urcorner$ solely on the basis of $\ulcorner \neg p \urcorner$, or solely on the basis of q, there is nothing one can do with the conclusion that one could not already do with the premise. Thomson illustrates this with the complex example of an oracle and an acolyte. The oracle sometimes contradicts itself, and then the acolyte has to figure out which statements to erase (Thomson 1990, p. 65). Suppose the oracle says q. It would be pointless for the acolyte to infer $\ulcorner p \rightarrow q \urcorner$, even though this follows logically. For

- If later the oracle said p, there'd be no need to do Modus Ponens to get q, because the acolyte already has q.
- If later the oracle says $\ulcorner \neg q \urcorner$, the acolyte couldn't use Modus Tollens[1] to get $\ulcorner \neg p \urcorner$. For, since his only basis for holding $\ulcorner p \rightarrow q \urcorner$ is his acceptance of q, on learning $\ulcorner \neg q \urcorner$ he would have to give up the conditional.

So there is no point to drawing the inference from q to $\ulcorner p \rightarrow q \urcorner$. That can explain our feeling that there is something wrong with this inference. But it doesn't give us grounds for thinking that the inference is invalid.

4.2 No truth conditions?

Recommended reading

Dorothy Edgington, "Do Conditionals Have Truth-Conditions?" (Edgington 1993).

Dorothy Edgington rejects the view that indicative conditionals are material conditionals, for reasons that go beyond those considered by Thomson. But, instead of proposing alternative truth conditions for indicative conditionals, she argues

[1] *Modus Tollens* is the inference from $\ulcorner p \rightarrow q \urcorner$ and $\ulcorner \neg q \urcorner$ to $\ulcorner \neg p \urcorner$.

that they do not have truth conditions at all. Conditionals, in her view, are not in the fact-stating business: they have a different role, which she seeks to explicate.

4.2.1 Arguments for the material conditional analysis

Edgington acknowledges that there are some powerful reasons for thinking that the material conditional analysis is right. Both of these inference forms seem good:

$$\text{Or-to-if}\ \frac{p \vee q}{\neg p \rightarrow q} \qquad \text{Not-and-to-if}\ \frac{\neg(p \wedge q)}{p \rightarrow \neg q}$$

Can you think of cases where you would be reluctant to make inferences with these forms?

However, if either of these inference forms are valid, $\ulcorner p \supset q \urcorner$ entails $\ulcorner p \rightarrow q \urcorner$. And we've already seen that if Modus Ponens is valid for '$\rightarrow$', $\ulcorner p \rightarrow q \urcorner$ must entail $\ulcorner p \supset q \urcorner$. So it looks like denying the equivalence of the indicative conditional and the material conditional requires either rejecting the validity of Modus Ponens, or rejecting the validity of both Or-to-if and Not-and-to-if.

Thomson has cautioned us to be skeptical about drawing conclusions about entailment from intuitions about the goodness of inferences. So there is room for maneuver here: we could try to explain why Or-to-if and Not-and-to-if inferences are good modes of reasoning *without* taking them to be valid. We'll see some examples of this strategy a bit later.

4.2.2 Arguments against the material conditional analysis

Partial acceptance

We have seen how Thomson defends the material conditional analysis against the most obvious objections by distinguishing between what is strictly speaking said and what is implied by the speaker's saying it. Edgington points out that this kind of story can at best explain why we refrain from *asserting* conditionals when we only know that their antecedents are false or their consequents true. She observes there are other data that cannot be explained in the same way: for example, data about the relation between our *degrees of confidence* in conditionals and our degrees of confidence in their antecedents and consequents.

Take an ordinary coin. What is your degree of confidence in (11)?

(11) If this coin is flipped, it will land heads.

Edgington says, plausibly, that you should have about 50% confidence in the conditional, assuming you think it's a fair coin.[2] Your degree of confidence in (11) presumably does not depend on how likely you think it is that the coin will be flipped. But the material conditional analysis predicts that it should! If the indicative conditional is a material conditional, then (11) will be true if the coin is not flipped, so you should get more confident that the conditional is true as you get less confident that the coin will be flipped. This, Edgington thinks, is a compelling reason to give up the view that indicative conditionals are material conditionals.

Thus Edgington concedes that Thomson's Gricean strategy works, if we confine ourselves to the conditionals we assert or accept with certainty. Her point is that it fails badly when we consider cases of uncertainty:

> This case against the truth-functional account cannot be made in terms of beliefs of which one is *certain*. Someone who is 100 percent certain that the Labour Party won't win has (on my account of the matter) no obvious use for an *indicative* conditional beginning 'If they win'. But someone who is, say, 90 percent certain that they won't win can have beliefs about what will be the case if they do. The truth-functional account has the immensely implausible consequence that such a person, if rational, is at least 90 per cent certain of any conditional with that antecedent. (Edgington 1993, p. 34)

Rejection

According to the material conditional analysis, *rejecting* a conditional requires accepting that its antecedent is true. But as Edgington notes, this seems wrong. For example, I might reject the conditional

(12) If the President sneezes tomorrow, the oceans will dry up.

without accepting that the President will sneeze tomorrow. I might, for example, think there's a 30% chance that the President will sneeze tomorrow, and also that there's no chance that the oceans will dry up tomorrow. In such a case, I would reject the conditional, but I would not accept the antecedent.

Bizarre validities

Edgington points out that the material implication account gives bizarre predictions about the validity of inferences. As one example, she gives William Hart's

[2] Do you agree? What is the alternative? One could simply take (11) to be false, when there is a chance that the coin will land tails, assigning the conditional a 0% degree of confidence. Against this, Edgington says: "if someone is told 'the probability is 0 that if you toss it it will land heads,' he will think it is a double-tailed or otherwise peculiar coin."

"new proof of the existence of God," which derives the existence of God from one's own failure to pray (Edgington 1993, p. 37):

1. If God does not exist, then it is not the case that if I pray my prayers will be answered. ($\neg G \rightarrow \neg(P \rightarrow A)$)
2. I do not pray. ($\neg P$)
3. Therefore (by the material conditional analysis), it is the case that if I pray my prayers will be answered. ($P \rightarrow A$)
4. So (Modus Tollens) God exists. (G)

She also notes that the material conditional analysis predicts that

$$(A \rightarrow B) \vee (\neg A \rightarrow B)$$

is a tautology. But intuitively it seems possible to reject both disjuncts. For example, if I know that Jack is on vacation in Bermuda, I might reject both 'If I go to the store, I will see Jack' and 'If I do not go to the store, I will see Jack'.

4.2.3 Rejecting Or-to-if

We noted that the inference pattern Or-to-if, together with Modus Ponens, would suffice to establish the material conditional view. Edgington does not offer a straightforward counterexample to Or-to-if: a case where we are certain of the truth of the disjunction and the falsity of the corresponding conditional. Instead, she deploys a principle linking entailment to degrees of confidence (Edgington 1993, p. 34):

(13) If A entails B, it is irrational to be more confident of A than of B.

Using this criterion, we can reject Or-to-if. Suppose you've rolled a die but you haven't seen how it landed. You think it's 1/6 likely that the die landed 1 and 1/6 likely that it landed 2. Now ask yourself:

(a) How likely is it that it landed either 1 or 2?

(b) How likely is it that, if it didn't land 1, it landed 2?

It seems rational to answer 1/3 to (a) and 1/5 to (b). But then, by our principle (13), the conditional 'if it didn't land 1, it landed 2' does not follow from the disjunction 'it landed either 1 or 2'.

If Or-to-if is invalid, why can't we find a straightforward counterexample? Edgington shows that in all *normal* cases where we would be prepared to assert a

disjunction '$A \vee B$', we should have high credence in the condition '$\neg A \rightarrow B$'. By normal cases where we would assert '$A \vee B$', she means cases where

- we have intermediate credence in both disjuncts, and
- we do not accept the disjunction on the basis of one of the disjuncts alone.

As she puts it: "If I am agnostic about A, and agnostic about B, but confident that A or B, I must believe that if not-A, B" (Edgington 1993, p. 40).

However, in cases where we have low credence in the disjunction (like the die case above), and cases where we accept the disjunction only because we think one of the disjuncts is very likely, Or-to-if fails rather obviously. Suppose you're 90% confident that it's 8 o'clock, but you think there's a small chance your clock is broken. Since you're 90% confident that it's 8 o'clock, you should be at least 90% confident that

(14) It is either 8 o'clock or 11 o'clock.

(It can't be rational to be less confident in a disjunction than in one of the disjuncts.) But you don't give high credence to the conditional

(15) If it is not 8 o'clock, it is 11 o'clock.

For, if it is not 8 o'clock (because the clock is broken), it is equally likely to be any other time.

In this way, Edgington explains both why Or-to-if seems so intuitively compelling, and why it is nonetheless invalid.

4.2.4 Edgington's positive view

So, what does Edgington think are the truth conditions of indicative conditionals, if they are not material conditionals? Her view is radical. She thinks that indicative conditionals do not have truth conditions at all. Conditionals are not "part of fact-stating discourse" (Edgington 1993, p. 46).

Instead saying under what conditions conditionals are true, Edgington proposes to explain their meanings by saying what mental states they express. When we judge that if A, B, she says, we are not judging that some proposition, *that if A, B*, is true. We are, rather, judging that *B under the supposition that A*. Similarly, when we judge it 60% likely that if A, B, we are not judging that some proposition (whose truth conditions we might try to articulate) is 60% likely to be true. Rather, we are judging that B is 60% likely to be true, under the supposition that A.[3]

The core of Edgington's positive view is the principle

[3] Her approach is a kind of *expressivism*, akin to Allan Gibbard's approach to normative language (Gibbard 2003), or Huw Price's approach to the language of probability (Price 1983).

> Conditional Likelihood X believes that (judges it likely that) if A, B to the extent that X judges that $A\&B$ is nearly as likely as A (Edgington 1993, p. 38).

If we represent judgments of likelihood as numerical probabilities, then this amounts to

> A person's degree of confidence in a conditional, if A, B, is the conditional probability he assigns to B given A.[4] (Edgington 1993, p. 39)

David Lewis (1976) showed that (given some plausible assumptions) there is no way to assign truth conditions to propositions of the form '$p \rightarrow q$' that will validate

> The Equation
> $$\Pr(p \rightarrow q) = \Pr(q|p)$$

So if Edgington is right that the degree to which you should believe $\ulcorner p \rightarrow q \urcorner$ is your subjective probability of p given q, then Lewis's triviality proof could be used as an argument for the no-truth-conditions view.[5] However, Edgington doesn't want to assume precise values, so she doesn't rely on the Lewis result. Instead she relies on an argument (discussed in the next section) that if the indicative conditional has truth conditions at all, it must be truth-functional. (She has already argued that the conditional is not truth-functional, so this suffices to establish the no-truth-conditions view.)

Because Edgington does not think that conditionals have truth values, she cannot think of validity as a matter of truth preservation. Instead, she embraces a notion of *probabilistic entailment* due to Ernest Adams.

> Probabilistic validity Let the *improbability*[6] of a proposition be 1 minus its probability. An argument is *probabilistically valid* just in case the improbability of the conclusion is guaranteed to be less than or equal to the sum of the improbabilities of the premises. (That is, for every probability function, the improbabilities of the premises sum to greater than or equal to the improbability of the conclusion.)

Note that a valid argument with many premises that have a high degree of probability can have a conclusion with a low degree of probability. For an example, consider

---

[4] The *conditional probability* of B given A is defined as follows (assuming $\Pr(A) > 0$):

$$\Pr(B|A) = \frac{\Pr(A \wedge B)}{\Pr(A)}.$$

[5] See Bennett 2003, ch. 5 for an accessible exposition and analysis of Lewis's argument.

[6] Adams uses the term "uncertainty" instead, but this has the odd result that propositions we are certain are false have the highest possible "uncertainty."

(16)
The die will not land on 1. [5/6 likely]
The die will not land on 2. [5/6 likely]
The die will not land on 3. [5/6 likely]
The die will not land on 4. [5/6 likely]
The die will not land on 5. [5/6 likely]
The die will not land on 6. [5/6 likely]

The die will not land on 1–6. [0/6 likely]

Valid arguments will preserve certainty, but they need not preserve degree of uncertainty.

4.2.5 Against truth conditions

On Edgington's view, there is no way to assign truth conditions to an indicative conditional: "there is no proposition such that asserting *it* to be the case is equivalent to asserting that *B* is the case given the *supposition* that *A* is the case" (Edgington 1993, p. 30). We have seen why she rejects the material conditional account, which is the only plausible truth-functional account of the conditional. But as we saw in Chapter 3, it is possible to give truth conditions for non-truth-functional operators and connectives. Perhaps this can be done for the indicative conditional? To rule this out, Edgington argues that if the indicative conditional has truth conditions at all, it must be truth-functional (Edgington 1993, pp. 42–46). Since she has already argued that the indicative conditional is not truth-functional, this gives her a general argument that it lacks truth conditions.

The main premise of her argument is the Conditional Likelihood principle stated above. The argument takes the form of a "tetralemma" (like a dilemma, but with four "horns" or alternatives instead of two). Suppose truth-functionality fails. Then the truth value of a conditional is not entirely determined by the truth values of its antecedent and consequent. So at least one of the following cases must obtain:

TT Some conditionals with a true antecedent and a true consequent are true and some are false.

TF Some conditionals with a true antecedent and a false consequent are true and some are false.

FT Some conditionals with a false antecedent and a true consequent are true and some are false.

FF Some conditionals with a false antecedent and a false consequent are true and some are false.

Exercise 4.1: Material conditionals

1. ⋆Construct a slingshot argument for the conclusion that the indicative conditional is truth functional.

2. ⋆Is Edgington right that anyone who accepts truth conditions for 'if' that sometimes make a conditional with true antecedent and consequent true, and sometimes false, must accept C_1 (defined on this page)? (Would it be possible to give an account on which *certainty* that the antecedent and consequent were true would suffice for certainty in the conditional, but the mere truth of the antecedent and consequent would not suffice for the truth of the conditional?)

Case TF can be ruled out straightaway: assuming Modus Ponens is valid, a conditional with a true antecedent and a false consequent cannot be true. That leaves three cases. Edgington is going to argue that none of them is possible. That will show that truth functionality can't fail.

If Case TT can obtain, she argues, then

> C_1. Someone may be sure that *A* is true and sure that *B* is true, yet not have enough information to decide whether 'If *A*, *B*' is true; one may consistently be agnostic about the conditional while being sure that its components are true (as for '*A* before *B*').

However,

> C_1 is incompatible with our positive account [Conditional Likelihood]. Being certain that *A* and that *B*, a person must think *A*&*B* is just as likely as *A*. He is certain that *B* on the assumption that *A* is true. (Edgington 1993, p. 44)

So this possibility must be rejected. "Establishing that the antecedent and consequent are true is surely one incontrovertible way of verifying a conditional" (Edgington 1993, p. 44).

The arguments against Case FT and Case FF rely on similar reasoning. For Case FT, Edgington argues that someone who is certain that *B* will have to regard *A*&*B* as just as likely as *A*, and by Conditional Likelihood this is sufficient for being certain that if *A*, *B*. Similarly, for Case FF, Edgington argues that if someone who knows that *A* and *B* have the same truth value (as would be the case if both were false) also knows that *A*&*B* is just as likely as *A*, and hence that if *A*, then *B*.

Notice that all of these arguments move from the observation that we would be certain that if *A*, *B* if we were certain about the truth values of *A* and/or *B*, to

the conclusion that it would be *true* that if A, B if A and/or B had certain truth values. Are these transitions warranted?

4.3 Stalnaker's semantics and pragmatics

Recommended reading

Robert Stalnaker, "Indicative Conditionals" (Stalnaker 1975).

Stalnaker agrees with Edgington that the material conditional analysis must be rejected, and that Or-to-if is invalid. But he gives a different sort of positive view, one that assigns truth-conditions to indicative and subjunctive conditionals in a modal framework.

4.3.1 Propositions, assertion, and the common ground

Before we look at Stalnaker's theory of conditionals, we need to sketch the theoretical background within which he gives his analysis.

The point of inquiry, as Stalnaker conceives it, is to distinguish between alternative ways the world could be. We start out in a state of ignorance. As far as we know, there are many open possibilities: the world could be this way or that way. A *possible world* is a maximally specific way the world might be: one that settles every question you could pose about the state of the world. As we inquire, we rule out possible worlds. The more opinionated we become about how things are, the fewer open possibilities remain.

A *proposition* is the content of a belief or assertion. When you believe that snow is white, for example, the thing you believe, namely *that snow is white*, is a proposition. For many purposes we can model propositions as functions from possible worlds to truth values: a proposition has the value *true* on the worlds in which it is true, and *false* on the worlds in which it is false. Equivalently, we can think of a proposition as a set of possible worlds: those at which it is true.

In this model, accepting a proposition is accepting that the actual world is a member of it. Rejecting a proposition is denying that the actual world is a member of it. And regarding a proposition as an open possibility is thinking that the actual world might be a member of it.

Assertion is a speech act that has its place in a *shared* process of inquiry. In any conversation, there is a *common ground* of propositions that are accepted within that conversation. (The participants may not actually believe these propositions, since one can accept a proposition, in the framework of a conversation, without believing it.)

The common ground is *common* in the sense that there is common knowledge about what is mutually accepted. If I am in doubt about whether you accept p, then p is not part of the common ground, even if in fact we all do accept it. Indeed, even if we all accept p, and we all *know* that the others accept p, p will fail to be in the common ground if we suspect that the others might not know that we accept p. To *presuppose* that p is to take p to be part of the common ground.

We can think of the common ground as a set of propositions. But we can also think of it as a set of possible worlds: those that fall into the intersection of the propositions. This set, which Stalnaker calls the *context set*, contains all of the worlds that are compatible with what the conversational partners mutually accept. Everything else is "off the table" and ruled out, for purposes of the conversation.

Once an assertion is made and accepted, its content is added to the common ground. We intersect the context set with the asserted proposition, throwing away worlds that aren't compatible with what is asserted. So, as the conversation progresses and more assertions are made and accepted, the context set shrinks. (Remember, removing worlds from the context set corresponds to *adding* information: the more propositions are accepted, the fewer worlds remain open possibilities.)

A proposition p is *accepted* in the context if p is true at every world in the context set.

4.3.2 Semantics

With this background in place, we can turn to Stalnaker's views about conditionals. Stalnaker thinks the truth conditions are the same for indicative and subjunctive conditionals. The difference between them has to do with the different *presuppositions* they carry.

> The idea of the analysis is this: a conditional statement, if A, then B, is an assertion that the consequent is true, not necessarily in the world as it is, but in the world as it would be if the antecedent were true. (Stalnaker 1975, p. 274)

More formally:

> $f(p, w)$ is a *selection function* that picks out the "closest" or "most similar" possible world to w at which p is true.
>
> $\ulcorner p \rightarrow q \urcorner$ is true at w if q is true at $f(p, w)$. (If $f(p, w)$ is not defined because there is no world where p is true, then the conditional is vacuously true.) (Stalnaker 1975, p. 275).

Constraints on the selection function

The selection function picks out the "closest" or "most similar" world in which the antecedent is true. But what is meant, exactly, by "closest" or "most similar"? There is no fixed answer: "Relevant respects of similarity are determined by the context" (Stalnaker 1975, p. 275). However, we can articulate three important constraints on selections functions:

C1 p is true at $f(p, w)$. (p is true at the closest world at which p is true.)

C2 If p is true at w, $f(p, w)=w$. (No world is closer to w than w itself.)

C3 If w is in the context set, then $f(p, w)$ must, if possible, be within the context set: that is, "all worlds within the context set are closer to each other than any worlds outside it."

While C1 and C2 hold for both indicatives and subjunctives, C3 is specific to indicative conditionals. When I use an indicative, the selection function has to pick out a world inside the context set. That is, the hypothetical situation we are considering must be compatible with everything we are already assuming about the actual world. When I use a subjunctive conditional, by contrast, I'm signaling that C3 does not apply: the closest world where the antecedent is true may be outside of the context set. This difference explains why subjunctives, but not indicatives, can felicitously be used when the antecedent is assumed to be false:

(17) Granted, I have a car.
But if I didn't have a car, I'd take the bus.
??But if I don't have a car, I'll take the bus.

4.3.3 Reasonable but invalid inferences

On Stalnaker's theory, the inference from q to $\ulcorner p \rightarrow q \urcorner$ is invalid. To get a countermodel, just suppose that q is true at the actual world, but false at the closest p-world.

Similarly, the inference from $\ulcorner p \vee q \urcorner$ to $\ulcorner \neg p \rightarrow q \urcorner$ (Or-to-if) is invalid. Here is a countermodel:

Context set = $\{w_1, w_2, w_3\}$

p is true at w_1 only, q is true at w_2 only.

$f(\ulcorner \neg p \urcorner, w_1) = w_3$

In this model $\ulcorner p \vee q \urcorner$ is true at w_1, but $\ulcorner \neg p \rightarrow q \urcorner$ is not true at w_1.

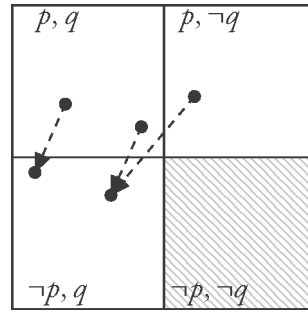


Figure 4.1: Or-to-if is a reasonable inference. The non-hatched rectangles represent the context set after $\ulcorner p \vee q \urcorner$ *has been appropriately asserted and accepted. No matter where we are in the context set, the closest world at which* $\ulcorner \neg p \urcorner$ *is true will be in the lower-left quadrant, and will therefore be a world where q is true. Thus* $\ulcorner \neg p \rightarrow q \urcorner$ *will be accepted as well. Note that the lower-left quadrant must be nonempty, since if it were empty the disjunction wouldn't have been "appropriately asserted."*

The inference from $\ulcorner \neg p \urcorner$ to $\ulcorner p \rightarrow q \urcorner$ and the inference from $\ulcorner \neg(p \wedge q) \urcorner$ to $\ulcorner p \rightarrow q \urcorner$ (Not-and-to-if) are also invalid. Of course, these inferences have to be invalid if we are to avoid the collapse of the indicative to the material conditional. But still, they can seem compelling, and we need to explain why.

Stalnaker calls an inference *reasonable* just in case, in any context where the premises "might appropriately be asserted," the conclusion will be accepted by a context if the premises are.[7] All valid inferences will be reasonable, in this sense, but some invalid inferences will also be reasonable. For an inference can be *acceptance-preserving* without being *truth-preserving*.

Or-to-if is a reasonable inference (see Fig. 4.1). Suppose $\ulcorner p \vee q \urcorner$ is appropriately asserted and accepted in the common ground C. We can now show that $\ulcorner \neg p \rightarrow q \urcorner$ must also be accepted in C. Let w be an arbitrary world in C. $\ulcorner \neg p \rightarrow q \urcorner$ is true at w just in case q is true at $w' = f(\ulcorner \neg p \urcorner, w)$. Since $\ulcorner p \vee q \urcorner$ was appropriately asserted, there must be at least one world in C at which $\ulcorner \neg p \urcorner$ is true, so by constraint C3,

[7] What is meant here by "might appropriately be asserted"? Instead of a general definition, Stalnaker offers necessary conditions that will suffice for the cases of interest to us here (Stalnaker 1975, pp. 277–8): "*It is appropriate to make an indicative conditional statement or supposition only in a context which is compatible with the antecedent. ...a disjunctive statement is appropriately made only in a context which allows either disjunct to be true without the other.*" Stalnaker clarifies the motivation for the latter condition: "If the context did not satisfy this condition, then the assertion of the disjunction would be equivalent to the assertion of one of the disjuncts alone. So the disjunctive assertion would be pointless, hence misleading, and therefore inappropriate."

$w' \in C$. Since $\ulcorner p \vee q \urcorner$ is accepted at C, $\ulcorner p \vee q \urcorner$ must be true at w'. By constraint C1, $\ulcorner \neg p \urcorner$ must be true at w'. So q must be true at w'. This suffices to show that $\ulcorner \neg p \rightarrow q \urcorner$ is true at w. Since w was an arbitrary world in C, it follows that $\ulcorner \neg p \rightarrow q \urcorner$ is accepted in C.

Thus, on Stalnaker's view, although the indicative conditional is not logically equivalent to a material conditional, it is

> ...equivalent in the following sense: in any context where either might appropriately be asserted, the one is accepted, or entailed by the context, if and only if the other is accepted, or entailed by the context. This equivalence explains the plausibility of the truth-functional analysis of indicative conditionals, but it does not justify that analysis since the two propositions coincide only in their assertion and acceptance conditions, and not in their truth-conditions. (Stalnaker 1975, p. 279)

In particular, as Stalnaker notes, the *denial* conditions for $A \vee B$ and $\neg A \rightarrow B$ are very different.

4.3.4 Contraposition and Hypothetical Syllogism

Stalnaker's semantics for conditionals makes Contraposition and Hypothetical Syllogism invalid:

$$\text{Contraposition} \quad \frac{\phi \rightarrow \psi}{\neg\psi \rightarrow \neg\phi} \qquad\qquad \text{Hypothetical Syllogism} \quad \frac{\phi \rightarrow \psi \quad \psi \rightarrow \xi}{\phi \rightarrow \xi}$$

It is easy to come up with counterexamples to these forms using subjunctives. Lewis 1973, p. 35 considers the following counterexample to Contraposition:

(18)
$$\frac{\text{If Boris had gone to the party, Olga would have gone.}}{\text{If Olga had not gone, Boris would not have gone.}}$$

Let us imagine that Boris stayed away from the party solely to avoid Olga, who was there. Olga, however, would have liked the party even better had Boris been there. In this scenario, the premise is true but the conclusion false.

Stalnaker gives the following counterexample to Hypothetical Syllogism with subjunctives (from Lewis 1973, p. 33):

Exercise 4.2: Stalnaker on conditionals

1. Show that Modus Ponens is valid for Stalnaker's conditional.
2. Give a countermodel to show that Contraposition is invalid on Stalnaker's semantics.
3. Give a countermodel to show that Hypothetical Syllogisim is invalid on Stalnaker's semantics.
4. Is Contraposition with indicative conditionals a *reasonable inference*, in Stalnaker's technical sense? Either show that it is not by giving an intuitive counterexample, or prove that it is.
5. Is Hypothetical Syllogism with indicative conditionals a *reasonable inference*, in Stalnaker's technical sense? Either show that it is not by giving an intuitive counterexample, or prove that it is.

(19)
If J. Edgar Hoover [the first director of the FBI] had been born a Russian, he would have been a communist.
If he had been a communist, he would have been a traitor.

So, if he had been born a Russian, he would have been a traitor.

Such counterexamples are possible because the "closest" relation is not transitive. (The closest yellow house to the closest blue house to me may not be the closest yellow house to me.)

Can you think of intuitive counterexamples to these inference forms with *indicative* conditionals? If not, does that cast doubt on Stalnaker's analysis?

4.3.5 The argument for fatalism

Stalnaker gives a beautiful application of his theory to an argument for fatalism discussed by Dummett (1964). Dummett imagines a civilian reasoning as follows during an air raid:

> Either I will be killed in this raid (K) or I will not be killed. Suppose that I will. Then even if I take precautions (P) I will be killed, so any precautions I take will be ineffective (Q). But suppose I am not going to be killed. Then I won't be killed even if I neglect all precautions; so, on this assumption, no precautions are necessary to avoid being killed (R). Either way, any precautions

> I take will be either ineffective or unnecessary, and so pointless. (Stalnaker 1975, p. 280)

The civilian decides not to take shelter and is killed. Clearly something has gone wrong in this reasoning, but what?

We can formalize the argument as follows:

1	$K \vee \neg K$
2	$\quad K$
3	$\quad P \rightarrow K$
4	$\quad Q$
5	$\quad Q \vee R$
6	$\quad \neg K$
7	$\quad \neg P \rightarrow \neg K$
8	$\quad R$
9	$\quad Q \vee R$
10	$Q \vee R$

Each step is plausible. What goes wrong?

On Stalnaker's view, the problems are in lines 3 and 7. The moves from 'K' to '$P \rightarrow K$', and from '$\neg K$' to '$\neg P \rightarrow \neg K$', ares not valid arguments, but merely reasonable inferences. When 'K' is accepted at a context, '$P \rightarrow K$' is accepted too. But when we're in a subproof, our hypothetical suppositions aren't accepted at our context. Remember, our context is compatible with both 'K' and '$\neg K$'. The feature we need arguments in subproofs to have is *truth preservation*, and this one isn't truth-preserving. "So it is a confusion of validity with reasonable inference on which the force of the argument rests" (Stalnaker 1975, p. 281).

4.4 Is Modus Ponens valid?

Recommended reading

Vann McGee, "A Counterexample to Modus Ponens" (McGee 1985).

Modus Ponens is often considered a paradigm of a valid inference. In nearly all discussions of the semantics of conditionals (including the preceding three

sections), it is taken for granted that Modus Ponens is valid. Vann McGee argues that this consensus is mistaken. Thinking about his argument will deepen our understanding of indicative conditionals.

4.4.1 The intuitive counterexamples

McGee begins by giving three intuitive counterexamples to Modus Ponens. In each case the major premise is a conditional with a conditional as its consequent.

The first counterexample concerns the 1980 US Presidential election, where Republican Ronald Reagan defeated Democrat Jimmy Carter. A third Republican candidate, John Anderson, ran as an independent and garnered a small fraction of the votes.

(20)
If a Republican will win the election, then if Reagan will not win, Anderson will win.
A Republican will win the election.

If Reagan will not win, Anderson will win.

It seems that the first premise was true, at the time of the election, because Anderson was the only other Republican candidate. And the second premise (we now know) was also true. But the conclusion was arguably false (at that context). After all, Anderson had virtually no chance of winning, and if Reagan hadn't won, Carter would have. Thinking of the conditional in the way Stalnaker recommends, the closest world to the actual world in which Reagan didn't win was a world where Carter won.

The second example concerns an animal seen from far away in a fishing net:

(21)
If that creature is a fish, then if it has lungs, it is a lungfish.
That creature is a fish.

If it has lungs, it is a lungfish.

The first premise is clearly true, because the only fish that have lungs are lungfish. And we may imagine that the second premise is also true (although we don't know this for sure). The conclusion, though, seems false. Lungfish are rare. If the creature in the net has lungs, it is very likely not a fish at all, but some other kind of sea animal.

The third example concerns poor Uncle Otto, who is digging a mine in his back yard, hoping to find gold or silver.

(22)
If Uncle Otto doesn't find gold, then if he strikes it rich, he will strike it rich by finding silver.
Uncle Otto won't find gold.

If Uncle Otto strikes it rich, he will strike it rich by finding silver.

The first premise is true, if we assume that gold and silver are the only minerals of value that could possibly be buried in the back yard. The second premise is also very likely true; it would be a huge coincidence if there were gold in Otto's back yard. But the conclusion seems false. Otto probably won't strike it rich at all, but if he does, it is just as likely to be by finding gold as by finding silver.

You might object that if we're certain of the premises of these arguments—certain, for example, that a Republican will win—then we must be certain of the conclusions (Katz 1999). Doesn't that show that the arguments are valid?

It does not. Edgington and Stalnaker have already given us several examples of inferences involving conditionals that are certainty-preserving, or acceptance-preserving, but not valid. For example:

(23) $\dfrac{q}{p \rightarrow q}$

On Edgington's account, if you are certain that q, you must be certain that if p, then q. Yet for Edgington, this inference is not valid: for, when q is not certain, it can be rational to have a lower confidence in the conditional than one has in q. On Stalnaker's account, if q is accepted in the common ground (and p is not ruled out by the common ground), $\ulcorner p \rightarrow q \urcorner$ must be accepted in the common ground too. Nonetheless, (23) is not valid. When it is not already common ground that q, it can be true that q even when the closest p-world is not a q-world.

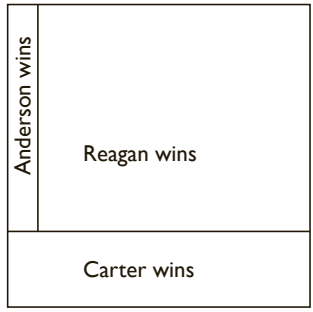
We can grant, then, that Modus Ponens is acceptance-preserving and certainty-preserving, while still raising a question about its validity. This question will have a different shape depending on whether you think of validity probabilistically (as Edgington does) or in terms of truth preservation (as Stalnaker does). So let us consider McGee's counterexamples from both points of view.

4.4.2 McGee's counterexamples as seen by Edgington

On Edgington's theory, the conclusion of a valid argument cannot have an improbability greater than the sum of the improbabilities of the premises. So, if we were to find an instance of an argument form where the two premises both have high probabilities (say, greater than 80%) and the conclusion a low probability

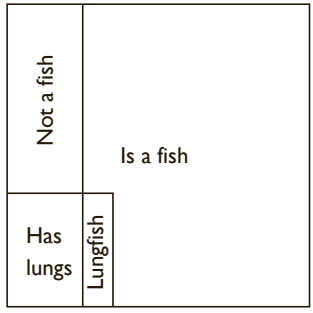
If a Republican will win the election, then if Reagan will not win, Anderson will win.
A Republican will win the election.

If Reagan will not win, Anderson will win.



If that creature is a fish, then if it has lungs, it is a lungfish.
That creature is a fish.

If it has lungs, it is a lungfish.



If Uncle Otto doesn't find gold, then if he strikes it rich, he will strike it rich by finding silver.
Uncle Otto won't find gold.

If Uncle Otto strikes it rich, he will strike it rich by finding silver.

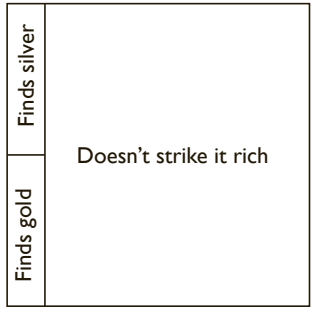


Figure 4.2: Credence diagrams for McGee's counterexamples. Bigger area = larger credence.

(say, less than 40%), that would count as a counterexample to the validity of the form.

With this in mind, examine the credence diagrams in Fig. 4.2. Here, the area occupied by a proposition represents the probability we give it. If we apply Edgington's Conditional Likelihood criterion for the acceptability of an indicative

conditional, we will see that in these cases the premises have high credences, but the conclusion has a low credence. For example, in the first counterexample, the first premise presumably has a credence of 1: after we conditionalize on a Republican's winning, the conditional probability that Anderson will win given that Reagan does not win is 1. The second premise also has a high credence: it is 75% likely that a Republican will win. But the conclusion has a very low credence. Conditional on Reagan not winning, Anderson is very unlikely to win. So we have a counterexample. The other cases have different structures, but they all generate counterexamples too, given Edgington's theory of validity.

4.4.3 McGee's counterexamples as seen by Stalnaker

Let us now consider the counterexamples in Stalnaker's framework. For Stalnaker, validity is truth preservation. So, a counterexample must have true premises and a false conclusion. Of course, on Stalnaker's view, indicative conditionals are context-sensitive: they can have different truth values, and express different propositions, at different contexts. So we need to keep the context fixed in evaluating the premises and the conclusion.

Let's suppose that, shortly before the 1980 election, Sarah utters the two premises and the conclusion of McGee's first counterexample. Suppose that the context set governing her conversation at this time includes worlds where Reagan will win, worlds where Carter will win, and worlds where Anderson will win.

Clearly, the proposition expressed in this context by the second premise of McGee's argument—that a Republican would win—was true at the actual world. The actual world, as we now know, is one in which Reagan would win.

Moreover, the proposition expressed by the conclusion—that at the closest world to the actual world at which Reagan won't win, Anderson will win—is false at the actual world. Worlds where Carter wins are much more similar to actuality than worlds where Anderson wins. (An Anderson victory would have required a miracle or a stunning October surprise.)

What about the first premise? It seems awfully hard to deny that, if a Republican wins, then if it's not Reagan it's Anderson. After all, Anderson and Reagan are the only Republicans in the race. But we know that Modus Ponens is valid for Stalnaker's conditional, so this first premise *can't* be true on his account. Let's see why it isn't.

On Stalnaker's semantics, we evaluate the first premise by, first, moving to the closest world to the actual world (@) where a Republican wins—call it w—and then evaluating the embedded conditional at that world. But since a Republican wins at the actual world, w is @! So the first premise is true at @ if the embedded conditional

(24) If Reagan will not win, then Anderson will win

is true at @. And, of course, it isn't: the closest world to @ at which Reagan doesn't win is a world at which Carter wins.

Thus, Stalnaker's semantics preserves the validity of Modus Ponens, but it does so at the cost of predicting that

(25) If a Republican will win the election, then if Reagan will not win, Anderson will win.

is false (at the envisioned context). This prediction seems wrong.

One reason it seems wrong is that (25) seems equivalent to

(26) If a Republican will win the election and Reagan will not win, Anderson will win.

Interestingly, (26) *is* true on Stalnaker's semantics. The closest world at which a Republican wins and Reagan doesn't is a world where Anderson wins. So Stalnaker's semantics opens up an unexpected gap between (25), which it takes to be false, and (26), which it takes to be true. The logical rule of

$$\textsf{Exportation}\ \frac{(p \wedge q) \rightarrow r}{p \rightarrow (q \rightarrow r)}$$

would allow us to infer (25) from (26). So, it seems, we have a counterexample to Exportation for Stalnaker's semantics. Saving Modus Ponens has a steep price.

4.4.4 Modus Ponens vs. Exportation

McGee shows that if we want a conditional that is stronger than the material conditional and weaker than logical implication, we need to choose between Modus Ponens and Exportation.[8] We can articulate the principle that the indicative conditional is weaker than logical implication thus:

StrImp If p logically implies q, then $\ulcorner p \rightarrow q \urcorner$ is true.

[8] The argument (McGee 1985, pp. 465–6) is similar to an argument from Gibbard 1981.

Simplifying a bit, the argument runs as follows:

1	$A \supset B$	Hyp
2	$(A \supset B) \wedge A$ logically implies B	(fact)
3	$((A \supset B) \wedge A) \rightarrow B$	StrImp 2
4	$(A \supset B) \rightarrow (A \rightarrow B)$	Exportation 3
5	$A \rightarrow B$	Modus Ponens for $\rightarrow$ 1, 4

So, if we have Exportation, Modus Ponens, and StrImp, the material conditional implies the indicative!

If we want to avoid this result, we need to give up one of these three principles. Giving up StrImp is unappealing: it means saying that a conditional could fail to be true even when the antecedent logically implies the consequent. So it's really a choice between giving up Modus Ponens and giving up Exportation. McGee argues that we should give up Modus Ponens, since there are intuitive counterexamples to Modus Ponens but not (he thinks) to Exportation. (Can you think of counterexamples to Exportation?)

Further readings

- Edgington 2014 and Bennett 2003 are useful surveys.
- Grice 1989 is an important resource for those who hope to defend the material conditional analysis of indicative conditionals. See also Rieger 2013, which summarizes a number of positive arguments for the material conditional analysis.
- On counterfactuals (not covered here), see Goodman 1955 and Lewis 1973.
- For more on the validity of Modus Ponens with indicative conditionals, see Kolodny and MacFarlane 2010.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5 Logical Consequence via Models

What is logic about? One common answer is that it is the science of good reasoning. This is too broad: there is more to reasoning well than being logical. Narrowing down a bit, we might say that logic is the science that tells us what follows from what. Its topic is the relation of logical consequence.[1]

But what is this relation? What do we mean, exactly, when we say that a sentence follows logically from some others (or equivalently, that an argument is logically valid)?

5.1 Informal characterizations of consequence

Of course, you've learned a technical definition of logical consequence, as preservation of truth in all models (§1.1.2). But it makes sense to ask for an informal characterization of the concept that this technical definition is trying to capture. After all, logicians were concerned with the systematization of valid arguments long before the concept of a *model* was articulated, and thus long before we could define logical consequence as preservation of truth in all models.

Historically, there have been three main approaches to characterizing consequence. We will consider them in turn, before turning to Tarski's classic article on the concept of logical consequence, which motivates and articulates the currently dominant model-theoretic conception.

5.1.1 In terms of necessity

Valid arguments are supposed to be truth-preserving: nothing untrue can be a logical consequence of true premises. Thus logical consequence would seem to involve at least

[1] Of course, logic also concerns itself with a number of other properties and relations: for example, validity, logical truth, logical equivalence, logical independence, consistency, and provability. But all of these have a close definitional or normative connection with logical consequence. (For example, p is logically independent of q and r if neither p nor $\ulcorner\neg p\urcorner$ is a logical consequence of q and r; and in a properly constructed proof system, only logical consequences of the premises should be provable.) So it is not unreasonable to think of logic as the science of logical consequence.

Material truth preservation An inference is *materially truth-preserving* iff it is not the case that its premises are true and its conclusion untrue.

However, material truth preservation does not seem to be sufficient for logical consequence.[2] The inference

(1) $\dfrac{\text{Wine is produced in France}}{\text{Wine is produced in Spain}}$

is materially truth-preserving (since both premise and conclusion are true), but not valid. Admittedly, learning that wine is produced in France might give one good reasons for thinking that wine is produced in Spain as well, if you already know that France and Spain are neighboring countries with similar climates and deep historical connections. Nonetheless, 'Wine is produced in Spain' does not follow *logically* from 'Wine is produced in France'.

What needs to be added to material truth preservation to get logical consequence? One common answer is *necessity*: in a valid argument, it is *impossible* for the premises to be true and the conclusion false. Logical consequence is thus a matter of

Necessary truth preservation An inference is *necessarily truth-preserving* iff necessarily, it is not the case that its premises are true and its conclusion untrue.

This way of thinking about logical consequence, which can still be found in many introductory logic textbooks, is widespread in the history of logic. We can trace it all the way back to Aristotle, who defined a *syllogismos* (valid deductive argument) as

> an argument in which, some things having been set down, something other than the things laid down follows by necessity through their being so. (*Prior Analytics*, 24b18–20)[3]

Necessary truth preservation suffices to rule out (1) as a valid argument. There is no *necessary* connection between the production of wine in France and the production of wine in Spain. One can imagine possible world-histories in which the growing of wine is strictly prohibited in Spain, but permitted in France. So, Necessary truth preservation does better than Material truth preservation is capturing our intuitive idea of "following from."

But recall from Chapter 3 that there are multiple ways of interpreting 'necessary'. When we say that something is necessary, we may mean that it is obligatory

[2] For an interesting argument that it isn't even necessary, see Field (2009b).

[3] Note that there is more to this definition than the appeal to necessity, though it is a matter of some dispute what is supposed to be added by "through their being so."

(deontic necessity), or that it could not have been otherwise (alethic necessity), or that it must be the case given what is known (epistemic necessity). So it is important to ask how the necessity in Necessary truth preservation is to be understood.

If we understand it as alethic necessity, then we must reckon with Kripke's observation that some necessary connections cannot be known a priori (§3.4.3). The inference

(2) $\dfrac{\text{Hesperus is visible}}{\text{Phosphorus is visible}}$

is necessarily truth-preserving, but it is not usually considered valid. One reason for this is that it is commonly assumed that the validity of inferences can be ascertained a priori. And no amount of armchair thinking about 'Hesperus is visible' and 'Phosphorus is visible' will show that one follows from the other.

Alternatively, we might take the necessity in logical consequence to be a kind of *epistemic* necessity. We might then say that an argument is valid just in case it can be known a priori that if the premises are true, so is the conclusion. On this account (2) is invalid. However, many inferences that are not traditionally considered logically valid will still be counted as valid in this sense. For example, the inference

(3) $\dfrac{\text{Boston is east of Seattle}}{\text{Seattle is west of Boston}}$

can be known a priori to be truth-preserving. And if mathematics is a priori, then all mathematical truths will count as logical truths (consequences of an empty set of premises). Some would take these consequences as reasons to reject Necessary truth preservation as an account of logical consequence. But defenders of Necessary truth preservation argue that logical systems are intended to capture only certain classes of valid arguments, so the fact that they do not count (3) as valid does not mean that (3) is not valid (Coffa 1975; Etchemendy 1990; Read 1994).

In arguing against the alethic construal of Necessary truth preservation, we appealed to the principle

> Apriority of logical consequence If a relation of logical consequence holds, it is knowable a priori that it is holds.

It is worth asking why one should accept this principle. One might think that the answer is this: the valid inferences are the ones we can rely on, the ones that are *guaranteed* not to take us from true premises to a false conclusion. If we could

not know a priori that the conclusion is true if the premises are, then we would not have this guarantee and we could not rely on the inference.

But on further consideration, this motivation starts to wobble. For, one might think, we can only rely on the guarantee a valid argument gives us when we *know* that conclusion is a logical consequence is the premises. But if we know that an argument is materially truth-preserving, then we already have a guarantee that the argument won't lead us from true premises to a false conclusion. We can rely just as much on materially truth-preserving inferences as on necessarily truth-preserving ones, as long as we *know* they are truth-preserving. This is why Bertrand Russell argues that the word 'implication' should be reserved for the relation of material truth preservation, and that features beyond this, like relations of form, are important only in that they make it possible to *know* that the relation of implication obtains (Russell 1920, p. 153).

There's another reason one might worry about characterizing logical consequence in terms of necessity or a priority. These notions are philosophically controversial. There is debate, for example, about whether anything can be known a priori at all, but those who deny this are not typically skeptics about the relation of logical consequence. Some philosophers have argued that there can be *empirical* reasons for rejecting the validity of certain inference forms of propositional logic (Quine 1951; Putnam 1968). This suggests that our basic understanding of logical consequence is not tied to any notion of necessity or apriority. How else might we characterize it?

5.1.2 In terms of proof

It would be natural to characterize logical consequence in terms of provability. After all, proofs are a canonical way of establishing relations of logical consequence. To say that a conclusion follows logically from some premises, we might try saying, is just to say that the conclusion can be proved from the premises taken as assumptions.

An obvious problem for this approach is that the notion of a proof seems to be system-relative. In a system containing a primitive rule for De Morgan's Laws, for example, this is a proof:

1	$\neg(S \vee T)$	Hyp	
2	$\neg S \wedge \neg T$	De Morgan's 1	(5.1)
3	$\neg S$	$\wedge$ Elim 2	

In our proof system (§1.1), by contrast, it is not a proof. One must instead do something like

1	$\neg(S \vee T)$	Hyp	
2	$\quad S$	Hyp	
3	$\quad S \vee T$	$\vee$ Intro 2	(5.2)
4	$\quad \bot$	$\neg$ Elim + Reit 1, 3	
5	$\neg S$	$\neg$ Intro 2–4	

Whether a sequence of formulas is a proof, then, is relative to a proof system—a set of rules specifying what counts as a proof. But whether a conclusion follows logically from some premises does not seem to be relative to a proof system.

Indeed, one way in which we evaluate proof systems is by asking whether they capture all and only the valid arguments (whether they are *sound* and *complete*, §1.1.5). We would regard it as a big strike against a proof system containing a rule like

Inverse Modus Ponens
$$\frac{p \supset q \quad q}{p}$$

that it allows us to construct proofs of invalid arguments. We seem to possess a notion of logical consequence, then, that is *not* relative to a specific system of proof rules.

Given that provability is system-relative, and logical consequence is not, is there any hope of characterizing logical consequence in terms of provability? We might simply pick a particular proof system. But this would make that proof system sound by stipulation. In addition, a specific proof system gives rules for proofs in a particular language. Its rules don't apply to proofs in a different language, even if the difference is a trivial one ('&' instead of '$\wedge$' for conjunction). But presumably arguments in any language can be classed as valid or invalid.

To avoid these problems, we might try saying that an argument is valid if it is provable in *some* system. This removes the system relativity—at the cost of making *all* arguments valid (since we can always come up with *some* system of rules that allows them to be proved). The obvious fix is to say that an argument is valid just in case it is provable in some *sound* proof system. But as a definition of validity, this is circular, since to say that a proof system is *sound* is to say that everything provable in it is *valid*.

In Chapter 6, we will look at an attempt by Dag Prawitz to solve these problems for proof-based definitions of validity.

5.1.3 In terms of counterexamples

We've considered the idea that we might define logical consequence in terms of proof—a natural idea, since we use proofs to establish that arguments are valid. But we might also look at it from the other side. How do we establish that arguments are *invalid*?

Not using proofs. One cannot show that an argument is invalid by proving the negation of its conclusion from its premises (see Exercise 1.2 (2)). Rather, we prove invalidity by giving a counterexample or a countermodel. The idea of a countermodel is later, and stems from Tarski's work, which we'll be looking at in §5.2. For now we will focus on the more intuitive idea of a counterexample (which goes back at least to Aristotle).

A *counterexample* to the validity of an argument is another argument with the *same form* that has true premises and a false conclusion. For example, suppose we are considering the argument

(4)
Most cats don't chase dogs.
Felix is a cat.

Felix doesn't chase dogs.

We can show that this argument is invalid by exhibiting another inference of the same form, which has clearly true premises and a clearly false conclusion:

(5)
Most people don't prove theorems.
Tarski is a person.

Tarski doesn't prove theorems.

In virtue of what do we say that (5) has the "same form" as (4)? Both are substitution instances of this schema:

(6)
Most *F*s don't *G*.
A is an *F*.

A doesn't *G*.

where *A*, *F*, and *G* are placeholders for expressions of a certain grammatical or semantic category (*A* a proper name, *F* a noun phrase, *G* a verb phrase).

You might reasonably ask, though, why (6) is "the" form of (4)? For (4) is also an instance of the schema

(7)
Most cats don't *G*.
A is a cat.

A doesn't *G*.

If *this* is the form of (4), then (5) does not have the same form and thus is not a counterexample. It is not enough to say that (6) is more general than (7), in the sense that every instance of the latter is an instance of the former but not vice versa. For the schema

(8)
P
Q

R

is even more general than (6), but we wouldn't regard an arbitrary instance of (8) with true premises and a false conclusion as a counterexample to (4).

The usual answer to this question is that only (6) gives the *logical form* of (4). Logical forms are supposed to consist entirely of logical words (like 'and', 'not', and 'all') and schematic letters. (7) does not give the logical form, because 'cat' is not a "logical word" or *logical constant*. And (8) does not give the logical form, because it deletes logical constants found in (4).

The logical form of an inference, then, is a schema derived from it by uniformly replacing the non-logical vocabulary with schematic letters and keeping the logical constants in place. But what is a logical constant, and why isn't 'cat' one?

One possible answer is that logical constancy, like provability, is a system-relative notion. Different systems treat different expressions as logical constants, and it is senseless to ask whether an expression is a logical constant *tout court*; we can only ask whether it is a logical constant in some system. In this case, the notion of logical form would be system-relative as well.

Many philosophers have sought to avoid this conclusion, arguing that certain expressions (by virtue of their meanings) are fit to be treated as logical constants, while others are not. A principled criterion for logical constants would make questions of logical form absolute, rather than system-relative. There has been considerable debate among philosophical logicians about whether there is such a criterion, and if so what it is, but this is not a topic we can pursue further here.[4]

If counterexamples are to be sufficient for showing invalidity, validity cannot consist in necessary truth preservation (on either an alethic or an epistemic understanding). Consider the argument

[4] For a survey of the main proposals for demarcating the logical constants, see MacFarlane 2017.

(9) Berkeley is north of Los Angeles.

Los Angeles is south of Berkeley.

This argument is necessarily truth-preserving on both the alethic and the epistemic understandings of necessity. Nonetheless, it is easy to give a counterexample (assuming that 'north' and 'south' aren't logical constants):

(10) Phoenix is south of Boston.

Boston is west of Phoenix.

(10) has the same logical form as (9) and is clearly invalid, since it has a true premise and a false conclusion. But why should *its* manifest invalidity be a reason to think that (9) is invalid, too? After all, (9) has a different premise and a different conclusion. The fact that there is no necessary connection between the premise and conclusion of (10) does nothing to show that there is no necessary connection between the premise and conclusion of (9). So, if validity is a matter of necessary truth preservation, counterexamples cannot prove invalidity.[5]

If we want a conception of validity on which counterexamples are sufficient for invalidity, we have three options. The simplest would be to understand validity as simply the absence of counterexamples. To say that an argument is valid is *just* to say that no argument with the same form has true premises and a false conclusion.[6] On this conception, it is clear why counterexamples demonstrate invalidity. On the other hand, the resulting notion of validity is very distant from the modal notion we discussed above. And it isn't so clear why it is an important notion. If validity and invalidity are properties of particular inferences, why should it matter whether *other*, formally similar inferences go from true premises to a false conclusion?

To avoid this worry, we might distinguish between validity, understood as necessary truth preservation, and a stronger notion of *formal validity*, understood as validity *in virtue of logical form*. On this conception, (9) can count as valid, despite the counterexample (10). What the counterexample shows is that the

[5] Despite this, it is common for elementary logic books to give a modal definition of validity and then proceed to use counterexamples to prove invalidity. Aristotle did the same thing in the *Prior Analytics*.

[6] Bernard Bolzano (1929/1931) gives the first clear statement of this idea. On Bolzano's account, to say an argument is valid (with respect to some selection of logical constants) is just to say that every argument obtained by uniformly substituting nonlogical expressions (preserving semantic categories) has either false premises or a true conclusion. Bolzano didn't think there was a principled answer to the question "what is a logical constant?" He thought you could consider different sets of different constants, and you'd get different logics. As we'll see in §5.2, Tarski's definition is a close cousin to Bolzano's.

validity of (9) is not due to its logical form, which it shares with (10). So the counterexample shows that (9) is not *formally valid*, not that it is not valid.[7]

Although this approach solves some of the problems with the first approach, it depends on some heavy conceptual resources. It requires us to make sense of both a modal notion of validity *and* the idea that the validity of some arguments is "due to" or "in virtue of" their forms. The first approach is more austere: it requires only the notion of logical form and generalization.

There is a third approach, which is even more austere than the first. It simply denies that counterexamples prove the invalidity of an argument. What a counterexample shows, instead, is that an argument *form* (or schema) is invalid. To say that an argument *form* is valid is to say that all of its instances are valid. Unlike the other approaches, it does not need to single out one of the argument's forms as "the" logical form, and as a result it does not need to demarcate the logical constants. We can see a single argument as having many forms. For example, the argument

(11) $\dfrac{\text{Some cats are mammals.}}{\text{Some mammals are cats.}}$

can be seen as having the (invalid) form

(12) $\dfrac{P}{Q}$

as well as the (valid) form

(13) $\dfrac{\text{Some } F\text{s are } G\text{s.}}{\text{Some } G\text{s are } F\text{s.}}$

Even (9) might be thought to have a valid form:

(14) $\dfrac{A \text{ is north of } B.}{B \text{ is south of } A.}$

[7] Notice that, on this approach, an argument that has no counterexamples could still fail to be formally valid: a counterexample shows that the validity of an argument *is not* explained by the form, but the absence of counterexamples need not show that the validity *is* explained by the form. In principle, it could be that all of the arguments that share a form are valid, but for reasons other than their possession of that form.

Knowing that an argument form is valid is useful, on this view, because it helps us to know that the particular arguments that are its instances are valid. But the use of counterexamples to establish invalidity has nothing to teach us about what validity for *arguments* amounts to (other than that it satisfies the minimal truth preservation criterion). We could keep the modal or epistemic definition for that, or we could take validity to consist in material truth preservation. More radically, we could argue that it is a category mistake to classify arguments (as opposed to argument forms) as valid or invalid.

5.2 Tarski's account of logical consequence

Recommended reading

Alfred Tarski, "On the Concept of Logical Consequence" (Tarski 1983a).

The modern notion of logical consequence was first articulated by Alfred Tarski (1983). Because Tarski's article has been so influential, we will go through it in detail, section by section. As you read Tarski, keep in mind the discussion of §5.1, and ask yourself what informal conception of logical consequence Tarski aims at capturing.

5.2.1 Tarski's aim

Tarski begins by stating his aim. He is not trying to define a new, technical concept. He wants to capture the ordinary concept of logical consequence—the one used by mathematicians when they say that one claim follows from another. However, he thinks, this ordinary concept is vague and people's usage differs, so there is no hope of capturing *every* feature that is part of anyone's intuitive notion of consequence: "We must reconcile ourselves from the start to the fact that every precise definition of this concept will show arbitrary features to a greater or less degree" (Tarski 1983a, p. 409).

5.2.2 Why proof-based approaches won't work

Tarski begins by considering whether logical consequence might be defined in terms of proof. He notes that this was the dominant view "until recently." Logicians had been convinced by their success in formalizing mathematical reasoning that they had isolated basic inference rules which "exhausted the content of the concept of consequence" (Tarski 1983a, p. 410).

Tarski objects to this idea—but not because of the system relativity of proof and provability (discussed in §5.1.2, above). He writes as if he would be content to

identify validity with provability in a system that was really capable of formalizing all mathematical reasoning. His objection is that no proof system can do this. No proof system can capture *all* of the valid inferences.

To show this, Tarski gives the following example of a valid inference that is not provable:

$$\begin{array}{ll} & A_0.\quad P(0) \\ & A_1.\quad P(1) \\ (15) & A_2.\quad P(2) \\ & \vdots\quad \text{(for all natural numbers)} \\ \hline & A.\quad \text{For all natural numbers } n, P(n) \end{array}$$

In order for A to be false, there would have to be a natural number n that n is not P. But for each n, there is a premise A_n that says that n is P. So, if the premises are true, the conclusion must be true. Tarski concludes that the conclusion "follows in the usual sense" from the premises (Tarski 1983a, p. 411). Yet the conclusion is not provable from the premises.

You might find the argument here puzzling. For our standard proof systems for first-order logic are known to be complete, relative to the standard semantics. That means that there are no valid inferences that cannot be proved. (15) is not an example of one, because (15) is not a valid inference of first-order logic. A first-order formalization of (15) would look like this:

$$\begin{array}{ll} & A_0.\quad Pa_0 \\ & A_1.\quad Pa_1 \\ (16) & A_2.\quad Pa_2 \\ & \vdots\quad \text{(for all natural numbers)} \\ \hline & A.\quad \forall x(Nx \supset Px) \end{array}$$

Here the predicate 'N' is a nonlogical constant, as are the constants 'a_0', 'a_1', 'a_2', ... So we can easily provide a first-order countermodel. For example, let the constants 'a_0', 'a_1', 'a_2', ...denote 0, 1, 2, ...; let the extension of 'N' be $\{-1, 0, 1, 2, 3, ...\}$; and let the extension of 'P' be $\{0, 1, 2, 3, ...\}$.

Some readers have taken Tarski to be assuming that 'N' and the numerals are logical constants, and so not up for reinterpretation in a model (Etchemendy 1990, p. 85). But this would make the argument dialectically weak: it would only persuade those who accepted an unorthodox view about the logical constants. A more plausible reading of this passage is that Tarski is assuming that we're working in a *type theory*: a higher-order logic, with variables ranging over objects, classes of

objects, classes of classes of objects, and so on.[8] In such a theory, one can define the numbers (and the concept *natural number*) in purely logical terms. For example, 0 is the class of all classes with no members, 1 is the class of all classes with exactly one member, 2 is the class of all classes with exactly two members, and so on. In a higher-order logic, then, (15) can be expressed using just the standard logical constants, but it cannot be proved using the standard rules.[9]

Could we simply add new rules that allow deriving A from A_0, A_1, ...? The most straightforward way would be to add a "rule of infinite induction," which allows one to infer A directly from the (infinitely many) A_is. To this Tarski objects:

> But this rule, on account of its infinitistic nature, is in essential respects different from the old rules. It can only be applied in the construction of a theory if we have first succeeded in proving infinitely many sentences of this theory—a state of affairs which is never realized in practice. (Tarski 1983a, p. 411)

One can get around this problem, however, if the logical language is powerful enough to express arithmetical concepts (as a type theory would be). For we can then construct a sentence B' that is true just in case A_0, A_1, etc. are all provable on the basis of the standard rules of inference.[10] We can now add a rule of inference that allows A to be inferred from B'—and we can keep adding rules of this kind.

No matter how many such rules we add, though, we can never capture *all* of the valid inferences. Gödel's first incompleteness theorem shows that higher-order logics are *essentially incomplete*: there is no hope of giving a complete finite set of axioms and inference rules for them. So even if we introduce a rule that allows (15) to be proven, there will inevitably be some other valid inference that cannot be proven. The upshot, Tarski thinks, is that "in order to obtain the proper concept of consequence, which is close in essentials to the common concept, we must resort to quite different methods and apply quite different conceptual apparatus in defining it" (Tarski 1983a, p. 413).

It is important to keep in mind that Tarksi's argument for this conclusion presupposes a higher-order logic. It would not be available if we restricted ourselves to first-order logic, which is complete. So, it is only because Tarski wants an analysis of consequence that is applicable to higher-order logics as well as first-order logic that he thinks proof-theoretic approaches are inadequate. (In §5.1.2, we gave other

[8] This is a generalization of second-order logic (§2.3), which only has the first two kinds of variables.

[9] In addition to making better sense of Tarski's argument, this reading is supported by the fact that Tarski uses a type theory in the paper he cites when introducing (15) (Gómez-Torrente 1996; Sagüillo 1997).

[10] Those who have studied Gödel's incompleteness results will see that this can be done using Gödel's technique of representing syntactical properties using arithmetical ones.

reasons earlier for thinking such approaches inadequate, even in the first-order case. But these sorts of considerations do not seem to have moved Tarski.)

5.2.3 Criteria of adequacy

Tarski says he wants to sketch a *general method* for defining logical consequence "for a comprehensive class of formalized languages." The basic idea, he says, is not original, but it has not been made explicit because the requisite semantic notions have not been available. But now that Tarski has given a rigorous definition of truth and satisfaction (Tarski 1935; Tarski 1983b), he can use these to state the idea rigorously.

He begins by giving two criteria of adequacy for a definition of consequence. The first is

> Truth Preservation If X is a logical consequence of a class K of sentences, then "it can never happen that both the class K consists only of true sentences and the sentence X is false."

There has been some debate in the literature about how to understand the scope of the modal "it can never happen" in this definition. One could read it with narrow scope:

(17) X is a logical consequence of $K \supset \neg\Diamond$(all K are true $\wedge$ X is false)

On this reading, Tarski is requiring that valid arguments be *necessarily* truth-preserving. But it is also possible to read the modal as having wide scope:

(18) $\neg\Diamond$(X is a logical consequence of $K \wedge$ all K are true $\wedge$ X is false)

On this reading, Tarski is requiring that valid arguments be *materially* truth-preserving.[11]

The second criterion of adequacy is

> Formality The relation of logical consequence "is to be uniquely determined by the form of the sentences between which it holds" and "cannot be affected by replacing the designations of the objects referred to in [the premises and conclusion] by the designations of any other objects" (Tarski 1983a, p. 415).[12]

This condition captures the idea that a valid argument is *counterexample-free* (§5.1.3). Instead of looking at the truth values of premises and conclusions in

[11] For the narrow-scope reading, see Etchemendy 1990, ch. 6 and Sher 1991, ch. 3. For the wide-scope reading, see Ray 1996. As we will see, a significant disadvantage of the narrow-scope reading is that some of Tarski's reasoning is only valid on the wide-scope reading.

[12] Note that in the type-theoretic framework Tarski is working with, the interpretations of predicates are also "objects" in the type hierarchy.

different possible worlds, we look at a whole class of arguments that share the "logical form" of our argument but have different "matter," that is, different nonlogical constants. If all of these are materially truth-preserving, the condition is satisfied.[13]

Tarski combines the two criteria into a condition he calls (F):

> (F) If, in the sentences of the class K and in the sentence X, the constants—apart from purely logical constants—are replaced by any other constants (like signs being everywhere replaced by like signs), and if we denote the class of sentences thus obtained from K by 'K'', and the sentence obtained from X by 'X'', then the sentence X' must be true provided only that all sentences of the class K' are true. (Tarski 1983a, p. 415)

(Note that the issue we raised about the scope of the modal in Truth Preservation arises for 'must' here as well.)

5.2.4 The insufficiency of (F)

Although Tarski takes (F) to be a necessary condition for logical consequence, he does not think that (F) is sufficient for logical consequence. His reason is that (F) may in some cases be satisfied "only because the language with which we are dealing does not possess a sufficient stock of extra-logical constants" (Tarski 1983a, pp. 415–16). To see why, suppose that the only terms in the language are 'Joe' and 'John' (both names of men) and the only predicates are 'is a man' and 'is a woman'. Now consider the inference

(19) $\dfrac{\text{Joe is a man}}{\text{John is a man.}}$

This inference is not valid. But it satisfies criterion (F), because there is no way to substitute terms for 'Joe' and 'John', and predicates for 'is a man', that gives you a true premise and a false conclusion. In a case like this, Tarski sees, an invalid inference can satisfy (F) just because the language isn't expressively rich enough to provide a counterexample.

We can put the concern this way: if we took (F) to be sufficient for validity, then whether an inference is valid would depend on the language's particular stock of nonlogical expressions. We could make an valid inference invalid just by adding to the language a word that doesn't occur in that inference. For example, if we added 'Sally', a name for a woman, to the language considered above, then (19) would not satisfy (F).

[13] The Formality criterion resembles Bolzano's substitution criterion, discussed in n. 6, above.

Tarski concludes that "(F) could be regarded as sufficient for the sentence X to follow from the class K only if the designations of all possible objects occurred in the language in question" (Tarski 1983a, p. 416). Not only is this not the case for existing languages; it arguably *can't* be achieved. Any language whose expressions are finite strings of symbols will contain at most countably many distinct terms.[14] But there are uncountably many real numbers. So inevitably there will be objects not denoted by any term in the language.

5.2.5 The semantic definition

Let's take stock. (F) tells us to look at a class of substitution instances of our inference, and see if any of them have true premises and a false conclusion. Whether they do depends on what their nonlogical expressions denote. The problem with taking (F) to be a sufficient condition for validity is that our language may lack expressions with the denotations needed to produce a counterexample. But, Tarksi sees, we can get around that problem if instead of substituting other nonlogical expressions for our nonlogical expressions, we substitute *variables*.

Suppose, for example, that our original inference is (19), above. Instead of thinking about a substitution instance like

(20) $\dfrac{\text{Joe is a man}}{\text{Sally is a man.}}$

where 'Sally' replaces 'John', we can think about

(21) $\dfrac{x \text{ is an } X}{y \text{ is an } X.}$

where the variables x, y, and X replace 'Joe', 'John', and 'man', respectively. We now ask whether there is any way of assigning values to the variables 'x', 'y', and 'X' that make the premises true and the conclusion false. In this case, there is such a way: for example, assign John to 'x', the set of men to 'X', and Sally to 'y'. And the existence of this assignment is independent of the language's stock of nonlogical expressions.

Where K is a class of sentences, let K' be the result of uniformly replacing nonlogical expressions of K with variables of the same semantic category. ("Uniformly" means that we replace the same nonlogical expression with the same

[14] To say that there are *countably many* terms is to say that the terms can be put in one-to-one correspondence with the natural numbers (or a subset of them).

variable, and different nonlogical expressions with different variables; "same semantic category" means that we use individual variables for names, class variables for predicates, and so on.) Tarski defines a *model* of K as an assignment of values to the variables that satisfies all the open formulas in K'.[15] He can then define logical consequence as follows:

> The sentence X *follows logically* from the sentences of the class K if and only if every model of the class K is also a model of the sentence X. (Tarski 1983a, p. 417)

This is a mathematically rigorous definition, because Tarski has shown elsewhere (Tarski 1935; see Tarski 1983b) how to give a rigorous definition of satisfaction, and hence of the *model of* relation.

When there are finitely many premises, logical consequence in the sense Tarski defines here reduces to the truth of a quantified sentence of higher-order logic. For example, (19) is valid if and only if

$$\forall X \forall x \forall y (Xx \supset Xy).$$

One can easily see that whether this condition holds does not depend on the language's particular stock of nonlogical expressions.

5.2.6 Satisfying the criteria of adequacy

Tarski says that it can be proved that his definition of logical consequence satisfies his criteria of adequacy, Formality and Truth Preservation, as embodied in condition (F). In evaluating this claim, we must return to the question whether the modal in Truth Preservation takes narrow scope (17) or wide scope (18).

On the narrow-scope interpretation, Truth Preservation says that if X is a logical consequence of K in Tarski's sense, then it cannot happen that: all the sentences in K are true and X is not true. It is not clear how this follows from Tarski's definition. Tarski's definition says that if X is a logical consequence of K, then there are no assignments of values to variables that satisfy a certain set of open formulas. That is a claim about what there is, not about what there must be. So if we read Truth Preservation with narrow scope, Tarski's definition does not vindicate it.

On the wide-scope interpretation, by contrast, Truth Preservation says that it cannot happen that: X is a logical consequence of K in Tarski's sense, all the

[15] Actually, instead of talking of assignments, he talks of sequences of objects. This is because an assignment of values to variables can be represented mathematically as a sequence: the first value is assigned to the first variable, the second to the second, and so on. Remember that "objects," for Tarski, includes items at all levels of the type hierarchy: ordinary objects, classes of objects, classes of classes, and so on.

sentences in K are true, and X is not true. This does follow straightforwardly from Tarski's definition.

Some commentators (Etchemendy 1990; Sher 1991; Sher 1996) argue that Tarski must have intended the narrow-scope reading, because only on this reading can he vindicate the ordinary, modal notion of logical consequence. These commentators bite the bullet and accuse Tarski of a modal fallacy. Others (Ray 1996) have argued that it is implausible that Tarski intended the narrow-scope reading, and even more implausible that he is guilty of a simple modal fallacy. On their view, Tarski is simply not interested in vindicating a modal conception of logical consequence.

5.2.7 Logical constants

Tarski's definition (like the contemporary model-theoretic definition) presupposes a division between logical and nonlogical terms of the language. To see this, consider a sentence like

(22) Cicero = Cicero

If 'Cicero' is the only nonlogical expression in this sentence, then a model of this sentence is an assignment that satisfies the open formula 'x=x'. Every assignment has that property so the sentence counts as logically true. But if we also regard '=' as a nonlogical expression, then a model of this sentence is an assignment that satisfies the open formula 'Xxy'. There are assignments that don't have that property, so the sentence doesn't count as logically true. Whether (22) is logically true, then—and whether it is a logical consequence of any class of sentences—depends on whether '=' is counted as a logical expression, or *logical constant*.

Going the other direction, suppose we counted 'Cicero' and 'Tully', as well as '=', as logical constants. Then

(23) Cicero = Tully

would count as a logical truth. (A model of it would be an assignment that satisfies 'Cicero = Tully'; since this sentence contains no free variables, every assignment satisfies it.) As Tarski notes:

> In the extreme case we could regard all terms of the language as logical. The concept of *formal* consequence would then coincide with that of *material* consequence.

That is, logical consequence would reduce to material truth preservation.

Although Tarski recognizes that certain choices of logical constants would "contradict ordinary usage," he says that he does not know any "objective grounds"

that would allow drawing a "sharp boundary" between logical and nonlogical terms (Tarski 1983a, p. 419). He concludes:

> Perhaps it will be possible to find important objective arguments which will enable us to justify the traditional boundary between logical and extra-logical expressions. But I also consider it quite possible that investigations will bring no positive results in this direction, so that we shall be compelled to regard such concepts as 'logical consequence', 'analytical statement', and 'tautology' as relative concepts which must, on each occasion, be related to a definite, although in greater or less degree arbitrary, division of terms into logical and extra-logical. (Tarski 1983a, p. 420)

Since the publication of Tarski's article, many philosophers have proposed objective criteria for distinguishing logical from nonlogical expressions.[16] Among them is Tarski himself, in a 1966 lecture published posthumously (Tarski 1986). However, one can also find Tarski voicing a more skeptical view, as in this 1944 letter to Morton White:[17]

> It is true, we can consider even the possibility of several non-equivalent definitions of 'logical terms' ...; e.g., sometimes it seems to me convenient to include mathematical terms, like the ϵ-relation, in the class of logical ones, and sometimes I prefer to restrict myself to terms of 'elementary logic'. Is any problem involved here? (White 1987, p. 29)

5.3 Interpretational and representational semantics

Let's take a step back and think abstractly about the definition of logical consequence as truth preservation at all models.

A model is a mathematical object of some kind. For Tarski, it is a sequence of objects from the type-theoretic hierarchy. In modern presentations of first-order logic, it is a pair consisting of a set of objects (the domain) and a function that assigns interpretations on that domain to the basic terms and predicates of a language. In truth-functional propositional logic, we could represent a model as an assignment of truth values to the atomic sentences. In a propositional modal logic, a model includes a set of worlds, an accessibility relation, a designated actual world, and a valuation function. The details of a model may differ from one logic to another, but in each case we define a function that maps a sentence (closed formula) of the language and a model to a truth value: the *truth-in-a-model* relation.

[16] For a survey, see Gómez-Torrente 2002 and MacFarlane 2017.

[17] As noted above (n. 6), Bolzano also took logical consequence to be a relative notion, relative to a selection of logical terms.

Our definition of logical consequence makes only minimal assumptions about models and truth-in-a-model. The models for a language can be any kind of mathematical object, as long as

- it is determined what counts as a model, so that we can quantify over all models, and
- it is determined which sentences are true at which models.

However, if we stay at this level of abstractness, it remains unclear what intuitive idea is represented by truth preservation in all models. For, as John Etchemendy has observed (1990), there are two very different ways we might think about what is represented by models and truth-in-a-model:

Representational semantics The models represent different ways the world might be—different possible situations. To say that a sentence is true in a model is then to say that this sentence, with its actual meaning, would be true in the situation represented by the model.

Interpretational semantics The models represent different possible meanings the sentences of our language might have had (different possible meanings for the language's nonlogical expressions). To say that a sentence is true in a model is then to say that this sentence, with a different interpretation of its nonlogical expressions, would be true given our actual situation.

These two conceptions are two different ways of thinking about the meaning or significance of models. Viewed representationally, models represent different ways the world could be, holding meanings of sentences fixed. Viewed interpretationally, they represent different meanings the sentences could have, holding the world fixed. These two conceptions of models can be linked to two of the informal characterizations of consequence we considered in §5.1, above. On the representational conception, truth preservation in all models amounts to necessary truth preservation (§5.1.1). On the interpretational conception, truth preservation in all models is a generalization of the idea of being counterexample-free (§5.1.3).

We saw how Tarski motivates his account as a generalization of condition (F), which articulates the idea of freedom from counterexamples. Tarski, then, seems to think of models in the interpretational way.

Indeed, it seems we must think of models this way or dramatically revise standard logical practice. For we standardly allow first-order models to assign any extensions at all to distinct atomic predicates. A model might then take the extensions of 'bachelor' and 'male' to be the set of even natural numbers and the set of odd natural numbers, respectively. It is easy to explain what this means in interpretational semantics: it represents an assignment of the meaning *even natural*

number to 'bachelor' and *odd natural number* to 'male'. But in representational semantics, it represents a situation, world, or state of affairs in which there are infinitely many bachelors and infinitely many males, but no overlap between these two groups. Is such a state of affairs really possible, in any sense? If not, then if we are doing representational semantics, we should not allow such a model.

Some philosophers have argued that the model-theoretic account seems intuitively plausible only because we fall back into thinking of the models representationally. If we view them interpretationally, as Tarski does, then as noted above, questions of logical consequence (for arguments with finitely many premises) boil down to questions about the truth of quantified sentences. The modal element that has often been thought to be part of the intuitive notion of consequence is nowhere to be seen. Dag Prawitz makes the criticism this way:

> It is said that with the help of valid inferences, we justify our beliefs and acquire knowledge. The modal character of a valid inference is essential here, and is commonly articulated by saying that a valid inference *guarantees* the truth of the conclusion, given the truth of the premises. It is because of this guarantee that a belief in the truth of the conclusion becomes justified when it has been inferred by the use of a valid inference from premises known to be true. But if the validity of an inference is equated with [a statement that the inference is counterexample-free] (or its variants), then in order to know that the inference is valid, we must *already* know, it seems, that the conclusion is true in case the premises are true. After all, according to this analysis, the validity of the inference just means that the conclusion is true in case the premises are, and that the same relation holds for all inferences of the same logical form as the given one. Hence, on this view, we cannot really say that we infer the truth of the conclusion by the use of a valid inference. It is, rather, the other way around: we can conclude that the inference is valid after having established for all inferences of the same form that the conclusion is true in all cases where the premises are. (Prawitz 2005, p. 675; for a very similar argument, see Etchemendy 1990, p. 93)

In the next chapter, we will consider Prawitz's own, very different way of defining consequence.

Further readings

- Etchemendy 1990 is an extended criticism of the Tarskian "interpretational" definition of logical consequence, and includes both exegesis of Tarski and conceptual argument. For a reply on Tarski's behalf, see Ray 1996. Etchemendy replies to some of his critics in Etchemendy 2008.

Exercise 5.1: Logical consequence

1. Show that it can be proved from Tarski's definition that if X is a logical consequence of a class K of sentences, then his condition (F) holds. What do you have to assume to prove this?
2. Compare Tarski's account of logical consequence with the modern model-theoretic account, as presented in §1.2.3. Which differences are superficial, and which differences are more significant? Can you think of any cases where the two accounts disagree about what follows logically from what?
3. How might one defend Tarksi's definition of consequence against Prawitz's criticism?

- On the problem of logical constants, see Gómez-Torrente 2002 and MacFarlane 2017.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

6 Logical Consequence via Proofs

In Chapter 5, we considered some problems with defining logical consequence in terms of proof:

- The notion of proof—and therefore provability—is system-relative, but logical consequence is not.
- If we try to get around this problem by defining consequence as provability in some *sound* system, then our definition is circular, since a sound system is defined as one in which only logical consequences can be proved.
- If we try to get around this problem by identifying a particular system, we face the problem posed by Gödel's incompleteness results: any proof system for second-order logic is incomplete, in the sense that there are logical consequences of a set of premises that cannot be proven from them in the system.

In this chapter, we will consider an approach to getting around these problems, worked out most fully by Dag Prawitz. A central idea is that inference rules of a certain form are *self-justifying*, so we can be confident that anything provable from them is a logical consequence.

After a look at the history of this idea, we will study Prawitz's proof-theoretic definition of consequence. We will see that this definition of consequence does not vindicate all classical logical consequences: it validates an alternative logic called *intuitionistic logic*. We will reflect on the nature of the dispute between classical and intuitionistic logicians—a paradigm case of fundamental logical disagreement.

6.1 Introduction rules as self-justifying

Recommended reading

Arthur N. Prior, "The Runabout Inference-Ticket" (Prior 1960).

Nuel D. Belnap, "Tonk, Plonk and Plink" (Belnap 1961).

6.1.1 Carnap's Copernican turn

The Preface to Rudolf Carnap's *The Logical Syntax of Language* contains this striking passage:

> Up to now, in constructing a language, the procedure has usually been, first to assign a meaning to the fundamental mathematico-logical symbols, and then to consider what sentences and inferences are seen to be logically correct in accordance with this meaning. Since the assignment of the meaning is expressed in words, and is, in consequence, inexact, no conclusion arrived at in this way can very well be otherwise than inexact and ambiguous. The connection will only become clear when approached from the opposite direction: let any postulates and any rules of inference be chosen arbitrarily; then this choice, whatever it may be, will determine what meaning is to be assigned to the fundamental logical symbols. By this method, also, the conflict between the divergent points of view on the problem of the foundations of mathematics disappears. For language, in its mathematical form, can be constructed according to the preferences of any one of the points of view represented; so that no question of justification arises at all, but only the question of the syntactical consequences to which one or other of the choices leads, including the question of non-contradiction. (Carnap 2002, p. xv)

Carnap is asking us to think about the relation between meaning and inference rules in a new way. The old way was to ask: given that '$\supset$' means 'if...then', what rules are valid for it? The question can seem hard to settle (as we saw in Chapter 4). Carnap is asking us to look at things the other way round, in what Coffa (1991) has described as a "Copernican turn" in semantics. Instead of trying to grasp an antecedent meaning and craft inference rules to it, Carnap proposes, we can just state some precise inference rules, and they will determine a meaning. Different rules will yield different meanings. Because the meaning of the constant is wholly constituted by the stipulated rules, these rules need no justification. It's just part of the meaning of '$\wedge$' that $\ulcorner \phi \wedge \psi \urcorner$ implies ϕ.

This is an appealing idea in many ways, and it became quite popular. Arthur Prior's classic paper "The Runabout Inference Ticket" (1960) is a tongue-in-cheek response.

6.1.2 Prior's article

Prior describes the view he is criticizing as holding that the meaning of 'and' can be given just by laying down introduction and elimination rules:

> Anyone who has learnt to perform these inferences knows the meaning of 'and', for there is simply nothing more *to* knowing the meaning of 'and' than being able to perform these inferences.

Knowing the meaning of 'and' does not require one to "grasp some concept" that goes beyond what ones gets just by learning these rules. One might naturally ask:

> How do we know that for any two statements P and Q there *is* a statement with the properties ascribed to "P & Q", i.e. a statement from which P and Q can both be derived, and which follows from P and Q together?

To this Prior replies:

> ...on the view we are considering such a doubt is quite misplaced, once we have introduced a word, say the word 'and', precisely in order to form a statement R with these properties from any pair of statements P and Q. The doubt reflects the old superstitious view that an expression must have some independently determined meaning before we can discover whether inferences involving it are valid or invalid. With analytically valid inferences this just isn't so.

The target view is plainly the one we have just seen in Carnap.

Prior now points out that, on these assumptions, we should be able to introduce a connective 'tonk' governed by these rules:

$$\text{tonk Intro}\ \frac{A}{A\ \text{tonk}\ B} \qquad\qquad \text{tonk Elim}\ \frac{A\ \text{tonk}\ B}{A}$$

We have stipulated that 'tonk' is the connective governed by these rules. So, on the view we are considering, these inferences are valid just in virtue of what 'tonk' means. But in a language with these rules, we can derive anything from anything:

1	A	Hyp
2	A tonk B	tonk Intro, 1
3	B	tonk Elim, 2

As Prior wryly notes, this new connective makes proving things extremely convenient!

6.1.3 Stevenson's response

Stevenson (1961) argued that the lesson of Prior's paper is that we need to take the semantic viewpoint. To introduce a connective, one must give a semantics (for example, a truth table). The rules are then justified in terms of the semantics. Carnap's idea that one can *define* a connective in terms of inference rules governing

it is just misguided. The problem with 'tonk' is that there doesn't exist any truth table that validates its introduction and elimination rules.

In addition to existence, Stevenson points out, there are issues of uniqueness. Suppose we define '%' as follows:

$$\%\ \text{Intro}\ \frac{A}{A\ \%\ B}$$

We haven't said enough to pick out exactly *which* connective '%' is. Any way of filling in the ?s in the following truth table will yield a connective for which % Intro is valid:

A	B	$A\ \%\ B$
T	T	T
T	F	T
F	T	?
F	F	?

We could make both ?s F, so that '$A\ \%\ B$' is equivalent to 'A', or we could fill them in as we do for disjunction, or we could make them both T. So in setting down % Intro we haven't yet said enough to pick out a unique connective.

Stevenson concludes that connectives must be introduced semantically, and the inference rules governing them must be justified in relation to the semantics. One must show that there is a meaning which validates the rules, and that there is not more than one such meaning. This amounts to the wholesale abandonment of Carnap's Copernican turn.

6.1.4 Belnap's Response

Nuel Belnap thinks this is an overreaction. Prior's example doesn't show that we have to give up the idea of defining connectives by their introduction and elimination rules, but only that we need to understand its limitations. It would be a shame, Belnap remarks, "to see the synthetic mode in logic pass away as a result of a severe attack of tonktitis" (Belnap 1961, p. 131).

So, what, according to Belnap, goes wrong with 'tonk'? The problem is that the rules for 'tonk' are inconsistent with assumptions we've already made. We assumed that our original, 'tonk'-free system gave the full story about deducibility for formulas not containing 'tonk'. In this system, we could not prove

(1) $A \vdash B$.[1]

[1] Here $\ulcorner \phi \vdash \psi \urcorner$ abbreviates $\ulcorner \psi$ is deducible from $\phi \urcorner$.

Exercise 6.1: Uniqueness of a connective

1. Check your understanding by convincing yourself that the rules for connective '%', described above, do not satisfy Belnap's uniqueness requirement. That is: if we had two connectives, '$\%_1$' and '$\%_2$', governed only by the introduction rule for '%', we could not prove

$$A \,\%_1\, B \vdash A \,\%_2\, B$$
$$A \,\%_2\, B \vdash A \,\%_1\, B$$

2. Show that the standard introduction and elimination rules for conjunction do satisfy Belnap's uniqueness requirement. That is, on the assumption that '$\wedge$' and '&' both satisfy these rules, prove that

$$A \wedge B \vdash A \,\&\, B$$
$$A \,\&\, B \vdash A \wedge B.$$

So we were committed to denying that B is deducible from A. But once we add 'tonk', we have to accept (1). Adding 'tonk' doesn't just give us new things to say about the deducibility of formulas containing 'tonk'; it forces us to take back our earlier claims about deducibility for formulas not containing 'tonk'. That is, the addition of 'tonk' fails to give us a conservative extension of our earlier system:

Conservative Extension A system S_c that adds rules for a new connective c to an existing system S is a *conservative extension* of S just in case for any statement ϕ not involving the new connective c, ϕ is provable in S_c only if ϕ is provable in S.

Belnap's appeal to non-conservativeness is a purely proof-theoretic way of capturing Stevenson's idea that "there is no such connective as 'tonk'."

Belnap also supplies a proof-theoretic way of capturing Stevenson's concern about *uniqueness*: To say that the rules for 'plonk' describe a unique connective is to say that if another connective 'plink' is given the same introduction and elimination rules, then 'plonk' and 'plink' are proof-theoretically equivalent:

$$A \text{ plonk } B \vdash A \text{ plink } B$$
$$A \text{ plink } B \vdash A \text{ plonk } B.$$

6.1.5 Prawitz's Response

Though he has defended the Carnapian idea that the meanings of connectives can be given by specifying introduction and elimination rules, Belnap has not vindicated the idea that the rules can be chosen arbitrarily and stand in no need of justification. On Belnap's view, a set of rules for a connective is justified only if it yields a conservative extension, and that is something that might require an external guarantee or proof.

Prawitz wants to defend the idea that introduction rules for connectives are *self-justifying*. He puts the point this way:

> For instance, if somebody asks why the rule for &-introduction...is a correct inference rule, one can answer only that this is just part of the meaning of conjunction: the meaning is determined partly by laying down that a conjunction is proved by proving both conjuncts, and partly by understanding that a proof of a conjunction could always be given in that way. (Prawitz 1985, p. 163)

Reiterating the idea in a later article, he says: "this amounts to making inferences by introduction valid—valid by definition, so to say" (Prawitz 2005, p. 694).

How, then, does Prawitz avoid tonktitis? By giving up the idea that *elimination* rules are self-justifying. On Prawitz's view, we can stipulate that a connective is governed by whatever introduction rules we like. Introduction rules must obey some formal constraints: they must introduce just one connective, and the premises must be subformulas of the introduced sentence. But other than that, anything goes. We need not worry about tonktitis, because we can't get a non-conservative extension just by adding introduction rules. (Convince yourself of this!)

But what about the elimination rules? Prawitz's approach is to show that they are valid by showing that anything that can be proved using elimination rules could, in principle, be proved without them. This process of eliminating elimination rules from a proof is called *normalization*.

So, as Prawitz sees it, the introduction rules are self-justifying, and the elimination rules are justified in terms of the self-justifying introduction rules. In the next section, we will see how Prawitz develops these ideas into a proof-theoretic account of logical consequence.

6.2 Prawitz's proof-theoretic account of consequence

Recommended reading

Dag Prawitz, "Logical Consequence From a Constructivist Point of View" (Prawitz 2005).

To motivate his account, Prawitz (1985) offers a helpful analogy. Consider these numerical expressions:

2	(4 + 16) − 17	the smallest multiple of both 4 and 3
1678	3 − 6	half of 15

Some of these expressions denote natural numbers, and others do not. How do we decide which do and which do not?

For '2' and '1678', the decision is trivial. They are in a "canonical form" for expressions denoting natural numbers: a sequence of digits. No substantive question arises about whether an expression of this form denotes a natural number. (For now, put aside metaphysical questions about whether *any* numbers exist.)

For the others, by contrast, a substantive question does arise. To determine whether these expressions denote natural numbers, one can see whether they can be reduced by arithmetical operations to canonical form. A bit of arithmetic reduces '(4 + 16) − 17' to '3', and that settles the question. By contrast, '3 − 6' and 'half of 15' cannot be reduced to the canonical form for natural numbers. So we can think of questions about whether expressions in canonical form denote natural numbers as trivial, and questions about whether other arithmetical expressions denote natural numbers as questions about whether these expressions can be reduced to a canonical form.

Prawitz's approach to the validity of arguments is similar. Arguments that consist in the application of an introduction rule are in canonical form, and no substantive question arises about their validity. They are valid in virtue of the meanings we have given their main connectives through stipulation of an introduction rule. Arguments that do not have this form—noncanonical arguments—are valid if they can be reduced, through a mechanical procedure, to canonical ones.

Already we can appreciate some advantages of this approach. First, because it works for *any* logical constants that can be defined via an introduction rule, our account of validity isn't tied to a particular proof system. Second, because the introduction rules are self-justifying, we do not need to stipulate that the system is a *sound* one (which, as we noted in §5.1.2, would be circular in an account of validity). Third, on this account, the validity of an argument consists in the fact that the decisions we have made in defining the logical constants commit us to

certain conclusions. In this way, Prawitz thinks, we can capture the necessary connection that has traditionally been thought to hold between premises and conclusion in a valid argument (see the quotation on p. 142, above).

Let's turn to the details.

6.2.1 Arguments

An *argument*, for Prawitz, is a step-by-step deduction, not a pair of premises and a conclusion. A *proof* is an abstract, non-linguistic entity that is "expressed" by a valid argument. (So a proof is to an argument as a proposition is to a sentence.)

To represent arguments, Prawitz uses Gerhard Gentzen's "tree-style" natural deductions, rather than the Fitch-style deductions we're used to. An argument deriving '$(A \wedge B) \wedge C$' from premises 'A', 'B', and 'C' looks like this:

$$(2) \qquad \dfrac{\dfrac{A \qquad B}{A \wedge B}\,\wedge\text{Intro} \qquad C}{(A \wedge B) \wedge C}\,\wedge\text{Intro}$$

The tree structure represents the dependence of steps on other steps. An argument deriving '$A \rightarrow B$' from a subproof of 'B' under the hypothesis 'A' looks like this:

$$(3) \qquad \dfrac{\begin{array}{c}[A] \\ \Delta \\ B\end{array}}{A \rightarrow B}\,\rightarrow\text{Intro}$$

The unasserted hypothesis 'A' is put in square brackets to indicate that it is discharged in the argument. 'Δ' serves as a placeholder for a valid argument from 'A' to 'B'.

Unbracketed premises are called *assumptions*, while bracketed premises are called *hypotheses*.

An argument is *closed* if it contains no assumptions and no unbound variables. Otherwise, it is *open*. (2), above, is open, because it has assumptions 'A', 'B', and 'C'. (3) is closed, because it has no assumptions or unbound variables.

6.2.2 Validity

What is it for an argument to be *valid*? That depends on whether it is open or closed:

- An *open* argument is valid iff replacing each assumption with a valid closed argument for that assumption always yields a valid argument.
- A *closed* argument is valid iff either

(a) it is a canonical argument, or
(b) there is an effective procedure for reducing it to a canonical argument for its conclusion.

- An argument is *canonical* if it ends with an introduction rule and contains valid (open or closed) arguments for the premises of the rule.[2]

An *effective procedure* is a mechanical method for solving a problem that is guaranteed to succeed after a finite number of steps (for example, the method you learned in school for doing long division).

For a non-canonical closed argument to be valid, there must be an effective procedure for transforming it into a canonical argument. This is most easily explained through some examples.

6.2.3 ∧ Intro and Elim

Let's first look at how we might show that the standard conjunction introduction and elimination rules are valid.

$$\frac{A \qquad B}{A \wedge B}\ \wedge \text{Intro} \qquad\qquad \frac{A \wedge B}{A}\ \wedge \text{Elim}$$

These are both open arguments, since they have assumptions. So, to show that they are valid is to show that the closed argument you get when you replace each assumption with a valid closed argument for that assumption is valid. Letting 'Δ_1', 'Δ_2', 'Δ_3' stand for valid closed arguments, then, we need to show that all instances of the following are valid:

(4) $$\frac{\begin{matrix}\Delta_1 & \Delta_2 \\ A & B\end{matrix}}{A \wedge B}\ \wedge \text{Intro}$$

(5) $$\frac{\begin{matrix}\Delta_3 \\ A \wedge B\end{matrix}}{A}\ \wedge \text{Elim}$$

[2] This definition looks circular, because canonical argument is defined in terms of valid argument, and valid argument is defined in terms of canonical argument. But the circle is not vicious. To settle whether an argument for ϕ is canonical, one need only settle the validity of arguments for formulas less complex than ϕ.

The astute reader will have noticed a gap in this account (which closely follows the presentation of Prawitz 2005). The notion of a canonical argument, and hence of a closed valid argument, is only defined here for formulas with introduction rules—and hence only for complex formulas. But there must also be valid closed arguments for atomic formulas, or every open argument with atomic formulas as premises would be (vacuously) valid. What counts as a closed valid argument for an atomic formula? Prawitz is clearer about this elsewhere, defining a notion of validity relative to a "system of canonical arguments for atomic sentences," and defining logical validity as validity relative to every such system (Prawitz 1985, p. 165; cf. Prawitz 2006, p. 515). For our purposes here, we can ignore this complication.

It is trivial to show that instances of (4) are valid. Any such instance will be a canonical argument, because it ends with the use of an introduction rule and contains valid arguments for the premises of the rule. For Prawitz, introduction rules need no further justification. They have the status of definitions of the logical constants they introduce.

More work is required to show that instances of (5) are valid, since they are not canonical arguments (they don't end with introduction rules). To show that such an instance is valid, we need to provide a method for transforming it into a canonical argument for its conclusion.

How do we do this? By assumption, Δ_3 is a valid closed argument for '$A \wedge B$'. So it must either be a canonical argument or be effectively reducible to one. A canonical argument for '$A \wedge B$' is one that ends in an application of the $\wedge$ Intro rule. So we can reduce Δ_3 to an argument of this form:

$$\dfrac{\begin{array}{cc}\Delta_4 & \Delta_5 \\ A & B\end{array}}{A \wedge B}\ \wedge \text{Intro}$$

But from this we can construct a canonical argument for 'A', the conclusion of the $\wedge$ Elim step. Because Δ_4 is valid and closed, it is either canonical or reducible to a canonical argument. Either way, we end up with a canonical argument for 'A'.

What have we done? We've shown that the $\wedge$ Elim rule is, in a sense, dispensable. Whenever there is a valid closed argument that uses $\wedge$ Elim, we can extract from it a valid closed argument that derives the same conclusion without using $\wedge$ Elim. In this way we can justify $\wedge$ Elim purely proof-theoretically.

6.2.4 $\vee$ Intro and Elim

Let us now see how this works for $\vee$ Intro and Elim:

$$\dfrac{A}{A \vee B}\ \vee \text{Intro} \qquad \dfrac{B}{A \vee B}\ \vee \text{Intro} \qquad \dfrac{\begin{array}{ccc} & [A] & [B] \\ A \vee B & C & C\end{array}}{C}\ \vee \text{Elim}$$

To show that $\vee$ Elim is valid, we need to show that any valid closed argument for its premise '$A \vee B$' can be reduced to a canonical argument for its conclusion 'C'. So, let Δ be a valid closed argument for '$A \vee B$':

$$\dfrac{\begin{array}{ccc}\Delta & [A] & [B] \\ A \vee B & C & C\end{array}}{C}\ \vee \text{Elim}$$

Since Δ is valid, it follows (from the definition of "valid closed argument" above) that Δ can be reduced to a canonical argument—that is, an argument for '$A \vee B$' that ends with an application of ∨ Intro and has valid arguments (and hence valid closed arguments) for the premises of the ∨ Intro step. We can get '$A \vee B$' using ∨ Intro either from 'A' or from 'B', so there are two possibilites here:

$$\dfrac{\begin{array}{ccc}\Delta_3 & [A] & [B] \\ A & \Delta_1 & \Delta_2 \\ A \vee B & C & C\end{array}}{C}\ \vee \text{Elim} \qquad\qquad \dfrac{\begin{array}{ccc}\Delta_4 & [A] & [B] \\ B & \Delta_1 & \Delta_2 \\ A \vee B & C & C\end{array}}{C}\ \vee \text{Elim}$$

But either way, we can rearrange our proofs to get a proof of 'C':

$$\begin{array}{c}\Delta_3 \\ A \\ \Delta_1 \\ C\end{array} \qquad\qquad \begin{array}{c}\Delta_4 \\ B \\ \Delta_2 \\ C\end{array}$$

Both of these are valid closed arguments for 'C', so (by the definition of valid closed argument) they can be reduced to canonical arguments for 'C', and we're done.

6.2.5 Philosophical reflections

What has Prawitz done? He has given an account of what it is for an argument to be valid that does not appeal to truth, satisfaction, or other semantic notions. It is also immune to Tarski's objections based on Gödel's incompleteness theorems, because it does not identify validity with derivability in any *particular* formal system. So Gödel's demonstration that no one formal system can capture all the valid arguments in a (sufficiently powerful) language need not bother us (Prawitz 1985, p. 166).

Prawitz's theory can be seen as an implementation of the idea that grasping the meaning of an expression is understanding how it is used. Of course, a logical constant like '∨' can be used in many different ways. However, Prawitz thinks, there is a *central* aspect of its use which determines all the others. This central aspect is captured in the canonical method for proving a disjunction: the ∨ Intro rule. It is in virtue of our grasp of this rule that we can appreciate the validity of other inferences involving '∨', such as ∨ Elim.

But how can we be sure that there *is* a central aspect of the use of '∨'? Prawitz's theory of meaning, and his definition of consequence, depends on a substantive assumption, which Dummett (1991, ch. 12) calls

Exercise 6.2: Prawitz's definition of consequence

1. Use Prawitz's method to show that Modus Ponens

$$\frac{A \qquad A \to B}{B} \to \text{Elim (Modus Ponens)}$$

   is a valid inference form, given the introduction rule for '$\to$':

$$\frac{\begin{array}{c}[A] \\ \Delta \\ B\end{array}}{A \to B} \to \text{Intro}$$

2. Explain why Prior's elimination rule for 'tonk' is not valid, according to Prawitz's definition of validity.

3. State an elimination rule that would make sense, given the introduction rule for 'tonk', and show that it is valid, according to Prawitz's definition.

The Fundamental Assumption: If a formula can be proved at all, a proof can always be given in canonical form.

If this assumption did not hold, then an argument might be valid even though it is not reducible to canonical form. But the Fundamental Assumption is by no means obviously true, even when applied to logical constants like '$\vee$'. Couldn't we have primitive proof rules that take us *directly* to '$A \vee B$' without passing through either disjunct? In his discussion, Dummett suggests that one might be able to tell that either a girl is in the garden or a boy is in the garden without being able to tell which disjunct is true. If that's right, then we can verify this disjunction in a way that cannot be reduced to an application of the disjunction introduction rule.

6.3 Intuitionistic logic

Prawitz's account of logical consequence is not only a philosophical alternative to Tarski's. It is also a *logical* alternative: it leads to a different answer to the question, "what follows from what?" Prawitz's criterion yields a logic called *intuitionistic logic*, which is strictly weaker than classical logic. Every intuitionist validity is a classical validity, but not vice versa.

In classical logic, the following argument is valid:

$$\frac{}{A \vee \neg A}$$

This is a closed argument, because it has no assumptions or free variables. So, on Prawitz's criterion, it is valid just in case it is (a) canonical or (b) reducible to a canonical argument. It isn't canonical, since a canonical argument for '$A \vee \neg A$' would have to proceed via $\vee$ Intro. So we must ask, is there an effective procedure for reducing it to a canonical argument?

Such an argument would look like one of the following:

$$\frac{\begin{array}{c}\Delta_1 \\ A\end{array}}{A \vee \neg A}\ \vee \text{Intro} \qquad\qquad \frac{\begin{array}{c}\Delta_2 \\ \neg A\end{array}}{A \vee \neg A}\ \vee \text{Intro}$$

But we have no ingredients in our original argument from which to construct an argument for either 'A' or '$\neg A$'. So there isn't going to be a way to reduce our argument to one of these. The argument isn't valid, then, on Prawitz's criterion. But it *is* valid on the classical criterion: every classical model makes '$A \vee \neg A$' true. So here is a real, logical difference between these two accounts of validity. Classical logic accepts the

Law of Excluded Middle $A \vee \neg A$,

while intuitionistic logic rejects it.

Another argument form that is classically, but not intuitionistically valid is

Double Negation Elimination $\dfrac{\neg\neg A}{A}$

The introduction rule for '$\neg$' is[3]

$$\frac{\begin{array}{c}[A] \\ \bot\end{array}}{\neg A}\ \neg \text{Intro}$$

Suppose we have a valid closed argument for '$\neg\neg A$'. Can we extract a canonical argument for 'A'? Reducing the valid closed argument for '$\neg\neg A$' to a canonical argument, we'd get:

$$\frac{\dfrac{\begin{array}{c}[\neg A] \\ \Delta \\ \bot\end{array}}{\neg\neg A}\ \neg \text{Intro}}{A}$$

[3] Prawitz simply defines '$\neg A$' as '$A \rightarrow \bot$', but this is the rule we'd get if '$\neg$' were treated as a primitive symbol.

But this doesn't give us the ingredients to construct a canonical argument for 'A'. So, Prawitz rejects Double Negation Elimination.

One might ask: why not take Double Negation Elimination *together with* the introduction rules as defining '¬'? But remember that the basis of Prawitz's reply to Prior is that only introduction rules can count as definitions of the connectives they introduce. Once we allow both introduction and elimination rules to count as definitions, we open ourselves to the possibility of tonkish connectives.

When we introduced the rules for '¬' in §1.1.3, we noted that the Double Negation Elimination rule was anomalous, in that it was the only rule that eliminated two connectives. From the intuitionistic point of view, it is an unjustifiable imposter that cannot be justified in terms of the introduction rule for '¬'. Removing this rule gives us a Fitch-style natural deduction system that is sound and complete for intuitionistic propositional logic.

Why the name "intuitionistic"?

Intuitionistic logic originated from the intuitionistic school in the philosophy of mathematics, which held that mathematics is about mental constructions (and thus "intuition" in roughly Kant's sense). To say that a mathematical object exists, on the intuitionistic conception, is to say that it is possible to construct it. Accordingly, one cannot demonstrate '$\exists x \neg Fx$' by deriving a contradiction from '$\forall x Fx$'. To prove '$\exists x \neg Fx$' one actually needs to construct an instance '$\neg Fa$', and it may be that, despite the provable absurdity of '$\forall x Fx$', no such instance is constructible. So intuitionistic mathematics refuses to allow certain forms of proof that are allowed in classical mathematics.

One can think of the intuitionists as identifying truth with provability—or, as Prawitz puts it, the "potential existence of evidence" (Prawitz 2005, p. 681). This makes a certain amount of sense in mathematics. Suppose there is a statement that is neither provable nor refutable. On the Platonist conception of mathematics, there is nonetheless a fact of the matter about whether it is true or false—a fact that is completely inaccessible to human investigators. For intuitionists, by contrast, as Michael Dummett writes,

> ...an understanding of a mathematical statement consists in the capacity to recognize a proof of it when presented with one; and the truth of such a statement can consist only in the existence of such a proof. (Dummett 1979, pp. 4–5)

If truth is identified with provability, and falsity with refutability, then acceptance of the Law of Excluded Middle amounts to acceptance of the claim that every statement is either provable or refutable. That is why intuitionists reject Excluded Middle.

6.4 Kripke semantics for intuitionistic logic

Prawitz's account is a *proof-theoretic semantics* for intuitionistic logic: it gives an account of validity based on provability. But in developing a feel for intuitionistic logic, and especially for proving that arguments are invalid, it is also useful to have a model-theoretic semantics. Here we present the basics of the "Kripke tree semantics" developed by Saul Kripke (1965). It is very similar to the possible-world semantics we have already seen for propositional modal logic.

A *Kripke model* for intuitionistic propositional logic is a quadruple $\langle W, \leq, @, V\rangle$, where

> W is a nonempty set of objects,
>
> $\leq$ is a *partial order* (a reflexive, transitive, and antisymmetric relation) on W,
>
> $@$ is a member of W, and
>
> V is a function (the *valuation*) that maps each element of W to a subset of the propositional constants, subject to the constraint that
>
> $$\text{if } w_1 \leq w_2, V(w_1) \subseteq V(w_2).$$

As with propositional modal logic, the triple $\langle W, \leq, @\rangle$ (without the valuation) is sometimes called a *frame*. The elements of W here can be thought of as nodes on a tree. The $\leq$ relation is an accessibility relation on the nodes. The valuation function assigns to each node the set of propositional constants that are true at that node. The assignment is required to be *persistent*: if a propositional constant is true at a node, it remains true as one ascends the tree. (One might think of the tree as representing a body of information or evidence; as time passes, more information is added, and information is never lost.)

We define truth at a model thus:

- If ϕ is a propositional constant, $\vDash^{w}_{\langle W,\leq,@,V\rangle} \phi$ iff $\phi \in V(w)$.
- $\vDash^{w}_{\langle W,\leq,@,V\rangle} \ulcorner \phi \wedge \psi \urcorner$ iff $\vDash^{w}_{\langle W,\leq,@,V\rangle} \phi$ and $\vDash^{w}_{\langle W,\leq,@,V\rangle} \psi$.
- $\vDash^{w}_{\langle W,\leq,@,V\rangle} \ulcorner \phi \vee \psi \urcorner$ iff $\vDash^{w}_{\langle W,\leq,@,V\rangle} \phi$ or $\vDash^{w}_{\langle W,\leq,@,V\rangle} \psi$.
- $\vDash^{w}_{\langle W,\leq,@,V\rangle} \ulcorner \neg\phi \urcorner$ iff there is no w' such that $w \leq w'$ and $\vDash^{w'}_{\langle W,\leq,@,V\rangle} \phi$.
- $\vDash^{w}_{\langle W,\leq,@,V\rangle} \ulcorner \phi \rightarrow \psi \urcorner$ iff for all w' such that $w \leq w'$ and $\vDash^{w'}_{\langle W,\leq,@,V\rangle} \phi$, $\vDash^{w'}_{\langle W,\leq,@,V\rangle} \psi$.

Conjunction and disjunction are treated extensionally: to see if a conjunction or disjunction is true at a node, one just has to look at whether its conjuncts or

disjuncts are true at that node. Negation and the conditional, by contrast, are intensional: the truth of a negation of conditional at a node depends on the truth of its constituents at other accessible nodes.

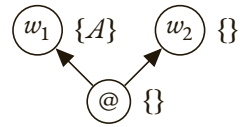
Given these clauses, the following fact can be established:

Persistence for Kripke models For any formula ϕ of intuitionistic propositional logic, if $\vDash^{w}_{\langle W,\leq,@,V\rangle} \phi$ and $w \leq w'$, then $\vDash^{w'}_{\langle W,\leq,@,V\rangle} \phi$.

We define truth at a model, logical truth, and validity just as we did in propositional modal logic:

> A formula ϕ is true in a model $\langle W, \leq, @, V\rangle$ if $\vDash^{@}_{\langle W,\leq,@,V\rangle} \phi$.
>
> A formula ϕ is logically true iff it is true in every Kripke model.
>
> An argument from premises Γ to conclusion ϕ is valid iff ϕ is true on every Kripke model on which all members of Γ are true.

Here's a model on which '$A \vee \neg A$' is not true:



In this model, 'A' is not true at @. But '$\neg A$' is not true at @ either, because 'A' is true at w_1. Given the clause for '$\vee$', it follows immediately that '$A \vee \neg A$' is not true at @.

We can show that the

Principle of Non-contradiction (PNC) $\neg(\phi \wedge \neg\phi)$

is intuitionistically valid by showing that there can be no Kripke model on which it fails to be true. For an instance of PNC to fail to be true on a Kripke model, we'd need $\ulcorner \phi \wedge \neg\phi \urcorner$ to be true at at least one node w accessible from @. This means that both ϕ and $\ulcorner \neg\phi \urcorner$ would have to be true at w. By Persistence, ϕ would have to be true at all nodes accessible from w. By the clause for '$\neg$', ϕ could not be true at any nodes accessible from w. Since $\leq$ is reflexive, w is accessible from w, so ϕ would have to be both true and not true at w—a contradiction. Thus every instance of PNC must be true on every Kripke model.[4]

[4] You might have the following worry about this reasoning: In assuming that ϕ cannot be both true and not true at w, aren't we just assuming the very principle whose truth is in question here, the Principle of Non-contradiction? How might one reply to this worry?

Exercise 6.3: Intuitionistic logic

1. Use Kripke models to establish the intuitionistic validity or invalidity of the following arguments:

   a) $\dfrac{A}{\neg\neg A}$

   b) $\dfrac{\neg\neg A}{A}$

   c) $\dfrac{\neg A}{A \rightarrow B}$

   d) $\dfrac{\neg\neg\neg A}{\neg A}$

   e) $\dfrac{A \rightarrow (B \vee C)}{(A \rightarrow B) \vee (A \rightarrow C)}$

   In which cases does the intuitionistic verdict depart from the classical one?

2. Do classicists and intuitionists disagree about the logical truth of any formulas whose only connectives are conditionals? If so, give an example. If not, explain why not.

3. Intuitionists accept the Principle of Non-contradiction, $\neg(\phi \wedge \neg\phi)$. How can they do this while rejecting the Law of Excluded Middle, $\phi \vee \neg\phi$? Aren't these two laws equivalent, given De Morgan's laws?

4. Prove Persistence for Kripke Models. *Hints:* Note that when ϕ is a propositional constant, the fact is guaranteed by the restriction on valuations that when $w \leq w'$, $V(w) \subseteq V(w')$. But it still needs to be proven that the fact holds for arbitrary (non-atomic) ϕ. To do this, you can argue by induction on the complexity of the formula. ϕ has the form '$\psi \vee \xi$, $\psi \wedge \xi$', '$\neg\psi$', or '$\psi \rightarrow \xi$'. So assume that the fact has already been established for ψ and ξ, and show that it holds as well for these compounds.

6.5 Fundamental logical disagreement

Recommended reading

Timothy Williamson, "Equivocation and Existence" (Williamson 1987), §1.

Intuitionistic logic and classical logic present themselves as rival theories of logical consequence. Classicists think that $\ulcorner \phi \vee \neg\phi \urcorner$ follows from anything, and that ϕ follows from $\ulcorner \neg\neg\phi \urcorner$. Intuitionists deny these claims. We have what appears to be a fundamental disagreement about basic questions of logic.

From the intuitionist's point of view, the classicist is a reckless reasoner, confidently drawing conclusions that do not follow by logic alone. As Anderson, Belnap and Dunn put it:

> The intuitionist overhearing with dismay the meanderings of some classicalist can always say: 'Poor fellow! He actually thinks he is reasoning. Still, there is some sense that can be made of his musings. What he seems to be doing is assuming (without warrant) a bunch of excluded middles. So I can charitably interpret him as constructing an enthymematic argument which can be made (intuitionistically) correct by adding the appropriate excluded middles as premises.' (Anderson, Belnap, and Dunn 1992, §80.4.1):

From the classicist's point of view, the intuitionist simply fails to acknowledge some obvious logical consequences. However, when called upon to vindicate the claim that Double Negation Elimination is valid, the classicist can offer only question-begging justifications. "Just look at the truth table," the classicist may say!

ϕ	$\neg\phi$	$\neg\neg\phi$
T	F	T
F	T	F

"Can't you see that ϕ and $\ulcorner \neg\neg\phi \urcorner$ have the same truth value on every row, and are therefore logically equivalent?" But the intuitionist can reply:

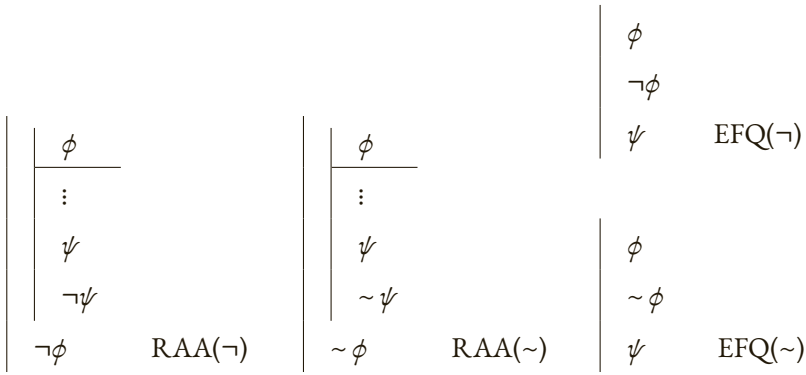
> You are assuming that ϕ is either true or false—otherwise the rows of your truth table wouldn't exhaust all the logical possibilities. But, given the equivalences
>
> ϕ is true iff ϕ
>
> ϕ is false iff $\neg\phi$,
>
> which we both accept, this is tantamount to assuming the Law of Excluded Middle. So, you are begging the question at issue.

6.5.1 Changing the subject?

It is tempting to suppose that this disagreement is really only a verbal one. If the meanings of the logical connectives are constituted by the rules that govern them, and the classicist and intuitionist associate different rules with '¬', then—one might think—they mean different things by '¬'. In that case, their dispute is merely verbal. Quine famously takes this line against those who countenance rejecting the Principle of Non-contradiction:

> They think they are talking about negation, '~', 'not'; but surely the notation ceased to be recognizable as negation when they took to regarding some conjunctions of the form '$p.{\sim}p$' as true, and stopped regarding such sentences as implying all others. Here, evidently, is the deviant logician's predicament: when he tries to deny the doctrine he only changes the subject. (Quine 1970, p. 81)

Are intuitionistic logicians "denying the doctrine," or just "changing the subject?" If the disagreement is merely verbal, then it ought to be possible to combine our different negation operators in a single language. (That is how we usually resolve merely verbal disputes: we distinguish the senses of the word and introduce explicit notation to disambiguate.) Let us use the symbol '~' for classical negation and '¬' for intuitionistic negation. Then, it might seem, both classicists and intuitionists can agree that Reductio Ad Absurdum (RAA) and Ex Falso Quodlibet (EFQ) are valid for both kinds of negation:



They should also be able to agree that Double Negation Elimination (DNE) is valid only for classical negation:

	$\sim\sim\phi$			$\neg\neg\phi$	
	ϕ	DNE(~)			~~DNE(¬)~~

However, things are not so simple. For, as Williamson shows, if we accept RAA(¬), RAA(~), EFQ(¬), and EFQ(~), then our two negations are provably equivalent:

1	$\neg\phi$	Hyp	1	$\sim\phi$	Hyp
2	$\quad\phi$	Hyp	2	$\quad\phi$	Hyp
3	$\quad\neg\phi$	Reit 1	3	$\quad\sim\phi$	Reit 1
4	$\quad\sim\phi$	EFQ(¬), 2, 3	4	$\quad\neg\phi$	EFQ(~), 2, 3
5	$\sim\phi$	RAA(~), 2–4	5	$\neg\phi$	RAA(¬), 2–4

It is an easy corollary that DNE(¬) must be valid if DNE(~) is (this is left as an exercise).

What this argument seems to show is that classicists and intuitionists can't regard themselves as talking about two different kinds of negation, defined by different rules. The intuitionist has to deny that there is any coherent meaning for a negation connective that would validate the classicist's rules. And the classicist, in turn, must deny that there is a coherent meaning for a negation connective that would validate RAA and EFQ but not DNE.

6.5.2 Interpreting classical logic in intuitionistic logic

Although there is no way for the intuitionist to make sense of classical negation as a general-purpose connective that can apply to any sentence, there is a way for the intuitionist to interpret all of the classicist's statements on which the classicist's reasoning is sound. This is the so-called *double-negation interpretation*. Define a function T from sentences of classical logic to sentences of intuitionistic logic as follows:

Double-negation interpretation

$$\begin{aligned} T(\phi) &= \neg\neg\phi \text{ (where } \phi \text{ is a propositional constant)} \\ T(\neg\phi) &= \neg T(\phi) \\ T(\phi \wedge \psi) &= T(\phi) \wedge T(\psi) \\ T(\phi \vee \psi) &= \neg(\neg T(\phi) \wedge \neg T(\psi)) \\ T(\phi \rightarrow \psi) &= \neg(T(\phi) \wedge \neg T(\psi)) \end{aligned}$$

It can be shown that an argument is classically valid just in case the argument that results from applying the T function to its premises and conclusion is intuitionistically valid.[5] If intuitionist uses T as a translation manual to interpret the things the classicist says, then the classicist will not be interpreted as endorsing any inferences the intuitionist regards as invalid. For example, when the classicist endorses

$$(6) \quad \frac{\neg\neg A}{A},$$

the intuitionist interprets this as

$$(7) \quad \frac{\neg\neg\neg\neg A}{\neg\neg A},$$

which is intuitionistically valid. Indeed, on this interpretation nothing the classicist can say can express what the intuitionist would mean by (6). From this point of view, the classicist is capable of expressing just *some* of the things the intuitionist can express, and as capable of appreciating just *some* of the inferences that are valid. As Burgess comments:

> while from one point of view intuitionistic logic is a part of classical logic, missing one axiom, from another point of view classical logic is a part of intuitionistic logic, missing two connectives, intuitionist '$\vee$' and '$\rightarrow$' (which are *not* intuitionistically equivalent to any compound involving just '$\neg$' and '$\wedge$'). From this point of view, what the classical logician accepts in accepting '$A \vee \neg A$' or '$\neg\neg A \rightarrow A$' is not what the intuitionist rejects in rejecting these same formulas. For the translations of '$A \vee \neg A$' and '$\neg\neg A \rightarrow A$' are just '$\neg(\neg A \wedge \neg\neg A)$' and '$\neg(\neg\neg A \wedge \neg A)$', both of which are perfectly acceptable intuitionistically. (Burgess 2009, pp. 129–30)[6]

[5] For a sketch of the proof, see Burgess 2009, pp. 127–9.

[6] Note Burgess has simplified his translations using the intuitionistic equivalence $\neg\neg\neg\phi \equiv \neg\phi$.

6.5.3 Interpreting intuitionistic logic in classical logic

The situation is, however, symmetrical: there is *also* a way for the classicist to translate the intuitionist's reasoning into classical logic—or to be precise, into classical logic extended with a modal operator.

From a classical point of view, it is natural to think that what intuitionists are really doing is talking about what is provable or verifiable. When intuitionists reject the Law of Excluded Middle, for instance, what they are really rejecting is the claim that every sentence is either provable or refutable. So interpreted, their claim is fully compatible with classical logic.

Using the modal '$\Box$' to mean *it is provable that*, we can construct the following translation T from the language of classical logic to the language of intuitionistic logic:

Modal interpretation

$$\begin{aligned} T(\phi) &= \Box\phi \text{ (where } \phi \text{ is a propositional constant)} \\ T(\neg\phi) &= \Box\neg T(\phi) \\ T(\phi \wedge \psi) &= T(\phi) \wedge T(\psi) \\ T(\phi \vee \psi) &= T(\phi) \vee T(\psi) \\ T(\phi \rightarrow \psi) &= \Box(T(\phi) \rightarrow T(\psi)) \end{aligned}$$

It can be proven that an argument is intuitionistically valid just in case its translation is valid in S4.[7] Interpreted using this translation manual, the intuitionist no longer disagrees with the classicist. When the intuitionist denies that 'A' follows from '$\neg\neg A$', for example, the classicist interprets her as denying that '$\Box A$' follows from '$\Box\neg\Box\neg\Box A$'. From this point of view, intuitionistic logic looks like a fragment of a classical modal logic. Everything that can be expressed in the intuitionistic language can be expressed in this classical language, but not vice versa.

Taking stock, then: it seems that we cannot regard the debate between the classicist and the intuitionist as merely a verbal one. Each party must reject the idea that there are coherent meanings for the connectives that yield the logic favored by the other. So each must deny that the other has got a *general logic*, one that is capable of expressing all of the logical distinctions. Yet each can interpret the other as speaking sensibly about just a *part* of logical space.

[7] For the proof, which relies on the close connections between Kripke models for intuitionistic logic and S4 models, see Burgess 2009, pp. 130–2.

Exercise 6.4: Intuitionistic and classical logic

1. Give an argument that if a system contains a classical conditional '⊃' that obeys our standard introduction and elimination rules, and an intuitionistic conditional '→' that obeys analogues of the same rules, then the conditionals are equivalent.
2. ⋆Can you think of any way to resist Williamson's argument that, in a system containing two negation connectives, both governed by RAA and EFQ, the negations will be equivalent?

6.5.4 Logical pluralism

So far we have not managed to vindicate Carnap's pluralistic conception of alternative logics. But there is a line of thought we have not yet explored that may come closer to that conception. Instead of seeing the classicist and the intuitionist as presenting compatible things to mean by 'negation', we might see them as presenting compatible things to mean by 'logical consequence'. According to this kind of pluralism (advocated in Beall and Restall 2006), each party could see the other as articulating a legitimate consequence relation and proving things about it. There would be no need to argue about which of these consequence relations was "the right one" or "the fundamental one." Both sides could agree that each consequence relation has its uses and its proper domains of application.

Williamson considers this way out but rejects it as "expressive of desperation rather than insight."

> As a matter of fact, both classical and intuitionistic logicians treat $X \vdash A$ as meaning that you are committed to A in making the set of assumptions X. It would otherwise be unclear that they could recognize each other as engaged in reasoning at all; to speak of classical and intuitionist logic would be to equivocate on the world 'logic'. Suppose that there were distinct but equally legitimate 'deducibility' relations, one classical and one intuitionist, and that you discovered your beliefs to have a certain consequence in the sense of one but not in the sense of the other; should you accept that consequence or not? Or if the classical 'deducibility' relation were applicable only to 'classical beliefs' and the intuitionist one only to 'intuitionist beliefs', which would a conjunction of a 'classical belief' and an 'intuitionist belief' be? (Williamson 1987, p. 112)

Williamson appeals here to some connections between logic and reasoning, and between logic and belief. Are these plausible connections, and is he right that they

render pluralism incoherent? Or is there a viable way to reject the idea that there is a single correct "general logic?" (We will circle back to some of these issues in §7.4.)

Further readings

- Heyting 1956, ch. 7 is the classic presentation of intuitionistic logic.
- Prawitz's earlier article (1985) is a somewhat more precise and technical presentation of his ideas.
- Michael Dummett is the most well-known defender of intuitionistic logic. Dummett 1979 is a technical examination of intuitionistic logic, but it also contains some excellent motivating philosophical discussion: see especially the Introductory Remarks and Chapters 1 and 7. Dummett 1978 gives an argument, from premises in the philosophy of language, for the conclusion that intuitionistic logic should be used not just in mathematics but everywhere.
- For more on logical pluralism, see Beall and Restall 2006, Field 2009a, Cook 2010, Shapiro 2014, and Steinberger 2019.

7 Relevance, Logic, and Reasoning

In classical logic, validity is typically understood in terms of truth preservation. A valid argument is one that can be counted on not to move from true premises to a false conclusion. The inference form

$$\text{Ex Falso Quodlibet}\ \frac{\phi \wedge \neg\phi}{\psi}$$

meets this description. For, as a matter of necessity, no matter how ϕ is interpreted, it is impossible for $\ulcorner \phi \wedge \neg\phi \urcorner$ to be true while ψ is false. This is impossible simply because it is impossible for $\ulcorner \phi \wedge \neg\phi \urcorner$ to be true. It doesn't matter what ψ is. Thus, anything follows logically from a contradiction—or so says the classical orthodoxy.

A minority tradition known as *relevance logic* has resisted this conclusion, on the grounds that the premises in a valid argument should be *relevant* to the conclusion and should not have an entirely different subject matter.[1]

In this chapter, we will look at some motivations for relevance logic and consider how a formal logic that excludes Ex Falso Quodlibet and other "irrelevant" consequences might be developed.

We will then consider the charge that the motivation for relevance logic stems from a confusion about the relation of logic to correct reasoning, and ask whether there is another way to motivate relevance logic.

[1] This idea was not always a minority one. According to Calvin Normore, "Ancient logics were all in some sense relevance logics. They insisted that for an argument to be valid, conditions must be met that guaranteed both that it would be impossible for the premises to be true and the conclusion false and that there would be connections of various kinds between the premises and conclusions" (Normore 1993, p. 448).

7.1 Motivations for relevance logic

Recommended reading

Robert Meyer, "Entailment" (Meyer 1971), pp. 812–18.

Robert Meyer gives (at least) five arguments for relevance logic.

Paradoxes of implication are counterintuitive Our logical theories must be judged against our intuitions about what follows from what, what is inconsistent, and so on. And our intuitions say that an arbitrary sentence does not follow from a contradiction:

> ...the claim that '$q\&{\sim}q$' entails 'p', in general, signals a breakdown in our intuitions not different in kind, though different perhaps in severity, from the kind of breakdown whose result is outright inconsistency...(Meyer 1971, p. 812)

Symmetry between overdetermination and underdetermination A theory may go wrong in either of two ways with respect to a pair of sentences ϕ, $\ulcorner\neg\phi\urcorner$. It may fail to tell us which is true (underdetermination), or it may tell us that both are true (overdetermination). Both of these are failures of discrimination; in both cases, we get no useful information about ϕ. So the cases should be treated similarly, but in classical logic there is a huge difference between them: overdetermination trivializes our theory—if the theory contains a single contradiction, it entails everything—while underdetermination is no big problem.

No reason to suppose mathematics is consistent There is no good reason to assume that mathematics must be consistent. If mathematics concerns a supersensible realm of objects, why should we assume they're like ordinary empirical objects with respect to consistency? If it is a free human creation, why can't it be inconsistent? However, classical logic *forces* a mathematical theory to be either consistent or trivial (meaning that it entails everything).

Contradictory beliefs People sometimes have contradictory beliefs. If they are committed to the logical consequences of their beliefs (as seems plausible), then according to classical logic, they are committed to everything. That is implausible. Even if your beliefs are inconsistent, we can meaningfully distinguish between the things you are committed to and the things you are not committed to.

Conflicting obligations Arguably, people are sometimes subject to conflicting obligations. As Lucy's friend, you might be obliged to tell her that her partner is thinking of leaving her, and also obliged not to tell her, because the information was given to you in confidence. Meyer suggests that "one is obligated to bring about what follows logically from what one is obligated to bring about." But then, if classical logic is the correct account of what follows logically, in cases of conflicting obligation one is obliged to bring about every state of affairs! That is not plausible.

7.2 The Lewis Argument

Recommended reading

John P. Burgess, "No Requirement of Relevance" (Burgess 2005).

Suppose we want a logic that rejects Ex Falso Quodlibet. What will it look like? It turns out that our options are severely constrained by a famous argument for Ex Falso Quodlibet which has come to be known as "the Lewis argument" (though it was known to Albert of Saxony and other medieval logicians centuries before it appeared in Lewis and Langford 1959, pp. 248–51). It runs as follows (here using 'P' and 'Q', but obviously generalizable to any formula):

1	$P \wedge \neg P$	
2	P	(1, $\wedge$ Elim)
3	$\neg P$	(1, $\wedge$ Elim)
4	$P \vee Q$	(2, $\vee$ Intro)
5	Q	(3, 4, Disjunctive Syllogism)

If we want to reject Ex Falso Quodlibet, we need to reject one of the assumptions of this argument. No one wants to reject $\wedge$ elimination. That leaves three good options:

- Reject $\vee$ Intro (Disjunctive Weakening)
- Reject the transitivity of entailment
- Reject Disjunctive Syllogism

Let us consider each of these in turn.

7.2.1 Rejecting Disjunctive Weakening

One might wonder how relevance can be represented formally. Intuitively, relevance often depends on one's background theory, and on the meanings of non-logical terms. Given what we know about the workings of matches, the fact that the match is wet is relevant to whether the match will light. But this is not the kind of relevance a logical theory can be sensitive to. The most a relevance logic can hope to do is codify *formal* relations of relevance.

One important formal surrogate for relevance is *variable sharing*.[2] In Ex Falso Quodlibet, a new propositional constant is introduced in the conclusion—one that is not shared by any of the premises. If this is why the premises of Ex Falso are not relevant to the conclusion, then Disjunctive Weakening is problematic for the same reason. In the move from 'A' to '$A \vee B$', we introduce a propositional constant 'B' not found in the premises.

W. T. Parry developed a relevance logic along these lines. In Parry's system (axiomatized in Parry 1933), ψ follows from ϕ only if all of the propositional constants in ϕ are in ψ. Disjunctive Weakening is rejected:

> a system might contain the proposition 'Two points determine a straight line', and yet not contain the proposition 'either two points determine a straight line or some angels have red wings'. In fact, a mathematician would rightly consider it, not only ridiculous, but utterly erroneous, to infer the latter proposition from the former. (Parry 1932, p. 119).

Parry called the resulting consequence relation *analytic implication*, alluding to Kant's definition of an analytic judgment as one in which the predicate concept is contained in the subject concept (Kant 1965, A6–7/B10).

Parry's logic of analytic implication blocks the Lewis argument. But it does so at a steep price. For example, it forces us to abandon the intuitively plausible contraposition rule for entailment:

Contraposition for Entailment If ϕ entails ψ, then $\neg\psi$ entails $\neg\phi$.

(To see why, let ϕ = '$A \wedge B$' and ψ = 'A'. Although '$A \wedge B$' analytically implies 'A', '$\neg A$' does not analytically imply '$\neg(A \wedge B)$', because the latter contains a constant not found in the former.) In addition, as Burgess (2005) observes, Disjunctive Weakening is an important tool in reasoning. (Consider how you would prove, for example, that every US President elected between 1950 and 2000 was either Democratic or Republican.) It is also hard to see why 'two points determine a straight line' should be irrelevant to the disjunction 'either two points determine a straight line or some angels have red wings'. What could be more relevant to the

[2] This is the name common in the literature, though it is more properly called *propositional constant sharing*.

truth of a disjunction than the truth of one of the disjuncts? Granted, it would be odd for a mathematician to draw this inference, but that can be explained without denying that it is valid: there is no mathematical point to drawing this inference.

For these reasons, the option of rejecting Disjunctive Weakening is not very popular among relevance logicians. (For further discussion, see Anderson and Belnap 1975, §29.6.1.)

7.2.2 Rejecting transitivity

Instead of focusing on the fact that Ex Falso Quodlibet introduces a new propositional constant in the conclusion, we might focus on the fact that it is truth-preserving only because it has a contradictory premise. The paradigm "fallacies of relevance" are classically valid inferences with contradictory premises or tautologous conclusions. So we might try to secure relevance by saying that an argument is valid just in case it is truth-preserving *and* has consistent premises and a non-tautologous conclusion.

As it stands, this is too crude. The following inferences would be excluded, even though intuitively there is no failure of relevance:

(1) a. $\dfrac{P \wedge \neg P}{\neg P}$ b. $\dfrac{P}{P \vee \neg P}$

G. H. von Wright (1957) and Peter Geach (1958; 1970) proposed that what distinguishes (1a) and (1b) from the irrelevant

(2) a. $\dfrac{P \wedge \neg P}{Q}$ b. $\dfrac{Q}{P \vee \neg P}$

is that there is a route to knowledge that (1a) and (1b) are truth-preserving that does not go via first establishing that the premise is contradictory or that the conclusion is tautologous. To know that (2a) and (2b) are truth-preserving, by contrast, one has to see that the premise of (2a) is necessarily false, and that the conclusion of (2b) is necessarily true.

Against von Wright's proposal, Strawson (1958) pointed out that there *is* a route to knowledge that (1a) and (1b) are truth-preserving that does not go via establishing that the premise is contradictory or the conclusion tautologous: one just has to construct a truth table and verify that there is no row where the premise is true and the conclusion false. In response, Smiley (1959) showed how

to implement von Wright's and Geach's basic idea more formally, without talk of knowledge. On Smiley's proposal (and that of Tennant 1994), an argument is valid iff it is a substitution instance of an argument that

- is classically valid,
- does not have a contradictory premise, and
- does not have a tautologous conclusion.

Thus, (1a) and (1b) are valid because they are substitution instances of classically valid arguments without contradictory premises or tautologous conclusions, namely:

(3) a. $\dfrac{P \wedge Q}{Q}$ b. $\dfrac{P}{P \vee Q}$

As you can verify, every step of the Lewis argument is valid in Smiley's sense. However, Ex Falso Quodlibet is not valid. So we have a case in which a sequence of valid steps produces an invalid argument:

- (1) entails (2)
- (1) entails (3)
- (2) entails (4)
- (3) and (4) together entail (5)
- But (1) does not entail (5)

Entailment, on this view, fails to be transitive!

The idea that entailment is non-transitive is a hard sell. Smiley himself regards his proposal as a way of formalizing the *intuitive* notion of entailment, but he thinks that a transitive notion should be used in logic, even if it has consequences, like the validity of Ex Falso Quodlibet, that seem unintuitive:

> the whole point of logic as an instrument, and the way in which it brings us new knowledge, lies in the contrast between the transitivity of 'entails' and the non-transitivity of 'obviously entails', and all this is lost if transitivity cannot be relied on. (Smiley 1959)

Nuel Belnap and Alan Ross Anderson put it more bluntly:

> Any criterion according to which entailment is not transitive, is *ipso facto* wrong. It seems in fact incredible that anyone should admit that B follows from A, and that C follows from B, but feel that some further argument was required to establish that A entails C. (Anderson and Belnap 1975, p. 154)

Indeed, the transitivity of entailment is simply assumed by natural deduction systems; if entailment were not transitive, then a deduction of a conclusion from a

premise would not show that the premise entails the conclusion. As Burgess (2005, p. 738) points out, there are formal proof systems (sequent calculi) which can work with non-transitive entailment relations, but giving up natural deduction seems a large cost. The practice of proving lemmas on the way to theorems in mathematics also presupposes that entailment is transitive. So, if entailment is not transitive, we cannot take mathematical practice at face value.

Even if we are willing to live with non-transitive entailment, there are reasons to doubt that Smiley's approach weeds out all of the "irrelevant" inferences. For example, consider

(4) a. $\dfrac{A}{A \wedge (B \vee \neg B)}$ b. $\dfrac{(A \wedge \neg A) \vee B}{B}$

These are classically valid inferences, the premises are not contradictory, and the conclusions are not tautologous. Yet they have the same odor of irrelevance as Ex Falso Quodlibet (Anderson and Belnap 1975, p. 155). (Consider how you would prove them.)

7.2.3 Rejecting Disjunctive Syllogism

There remains just one option for rejecting the Lewis argument: rejecting the use of

Disjunctive Syllogism $\dfrac{\phi \vee \psi \quad \neg\phi}{\psi}$

at step 5. This is, in fact, the most popular approach to developing a relevance logic.

Although Disjunctive Syllogism does not immediately strike one as "irrelevant," the way Ex Falso Quodlibet does, it seems more worthy of suspicion than Disjunctive Weakening. Disjunctive Weakening is the standard introduction rule for '$\vee$'. So if we accept Prawitz's view that the introduction rules define the connectives, there is no good basis for rejecting it. One can only argue about what elimination rules are valid, given this introduction rule (Dunn and Restall 2002). Disjunctive Syllogism, by contrast, is an elimination rule, so we can ask whether (like *tonk*) it lets us get out of a disjunction and a negation more than the introduction rules put in. J. Michael Dunn and Greg Restall argue that

> the problem with the Disjunctive Syllogism is just that $p \vee q$ might have been introduced into discourse (as it is in the Lewis 'proof') by $\vee$-Intro from p.

Exercise 7.1: Disjunctive Syllogism

How would you show that Disjunctive Syllogism is valid using the introduction and elimination rules for '$\vee$' and '$\neg$' from §1.1?

> So then to go on to infer q from $p \vee q$ and $\neg p$ by the disjunctive syllogism would be legitimate only if the inference from p, $\neg p$ to q were legitimate. But this is precisely the point at issue. At the very least the Lewis argument is circular (and not independent). (Dunn and Restall 2002, p. 35)

Although this is not itself a reason for everyone to reject disjunctive syllogism, it is a reason for those who reject the validity of Ex Falso Quodlibet to do so.

Note that Disjunctive Syllogism is essentially Modus Ponens for the material conditional, since $\ulcorner \phi \vee \psi \urcorner$ is equivalent to $\ulcorner \neg \phi \supset \psi \urcorner$. Relevantists do not think the material conditional is a real conditional at all, so giving up this principle is not a problem for them.

For a concrete example of a logic that rejects Disjunctive Syllogism, we'll focus on Belnap and Anderson's logic of first-degree entailment (called $\mathsf{E}_{\mathsf{fde}}$), which is equivalent to a logic devised by Ackermann (1956, pp. 113–128).

7.3 First-degree entailment

Recommended reading

Alan Ross Anderson and Nuel D. Belnap, Jr., *Entailment: The Logic of Relevance and Necessity* I (Anderson and Belnap 1975, pp. 150–166).

Because we can always combine (a finite number of) premises into a single conjunction, we'll consider here only one-premise arguments, which we'll call *entailments*. The goal will be to define a notion of *tautological entailment* that captures just the ones that are relevantly valid in virtue of their propositional forms. We'll see how this can be done in two different ways: first, through a syntactic decision procedure, and then using four-valued truth tables.

7.3.1 A syntactic procedure

First, some definitions. A *atom* is a propositional constant or its negation:

atoms	not atoms
P	$P \vee Q$
$\neg P$	$P \wedge \neg R$

A *primitive conjunction* is a conjunction of atoms. A *primitive disjunction* is a disjunction of atoms. A single atom counts as both a primitive conjunction and a primitive disjunction.

primitive conjunctions	primitive disjunctions	neither
$P \wedge \neg P \wedge Q$	$P \vee Q \vee \neg P$	$(P \wedge Q) \vee R$
$P \wedge Q \wedge R \wedge \neg S$	$P \vee R$	$Q \wedge (P \vee \neg R)$
P	P	$\neg\neg P$
$\neg P$	$\neg P$	$P \vee \neg\neg Q$

$\ulcorner \phi \Rightarrow \psi \urcorner$ is a *primitive entailment* if ϕ is a primitive conjunction and ψ a primitive disjunction. (The terminology is a bit confusing, because ϕ need not actually entail ψ. A better term would be 'primitive entailment statement'.)

primitive entailment	not a primitive entailment
$P \wedge \neg P \wedge Q \Rightarrow P \vee R$	$P \wedge \neg P \wedge Q \Rightarrow P \vee (R \wedge S)$
$P \Rightarrow R$	$\neg P \Rightarrow P \wedge R$

A primitive entailment $\ulcorner \phi \Rightarrow \psi \urcorner$ is *explicitly tautological* if some (conjoined) atom of ϕ is identical with some (disjoined) atom of ψ.

explicitly tautological	not explicitly tautological
$P \wedge \neg P \wedge Q \Rightarrow P \vee R$	$P \wedge \neg P \Rightarrow Q$
$\neg P \wedge \neg Q \wedge \neg R \Rightarrow S \vee \neg Q$	$Q \Rightarrow P$

This captures a certain kind of "containment" of conclusion in premises.

Entailments that are *not* primitive entailments, like '$P \vee Q \Rightarrow P \vee \neg(R \wedge \neg R)$', are tested using the following procedure:

1. Put the premise into *disjunctive normal form* (a disjunction of primitive conjunctions, or a single primitive conjunction) and the conclusion into *conjunctive normal form* (a conjunction of primitive disjunctions, or a single primitive disjunction). You should now have something of the form

$$\phi_1 \vee \phi_2 \vee \ldots \vee \phi_n \Rightarrow \psi_1 \wedge \psi_2 \wedge \ldots \wedge \psi_n,$$

where each ϕ_i is a primitive conjunction and each ψ_i is a primitive disjunction.

2. The entailment is a *tautological entailment* iff for every ϕ_i and ψ_j, $\ulcorner \phi_i \Rightarrow \psi_j \urcorner$ is explicitly tautological.

This procedure depends on the fact that every entailment can be put into normal form using relevantly acceptable rules. Here is the algorithm:

a) Apply De Morgan's laws and Double Negation Elimination to drive negations inward as far as possible.

De Morgan's Laws

$$\neg(\phi \vee \psi) \iff \neg\phi \wedge \neg\psi$$
$$\neg(\phi \vee \psi) \iff \neg\phi \wedge \neg\psi$$

Double Negation Elimination

$$\neg\neg\phi \iff \phi$$

(Note that these are substitution rules (§1.2.4); one can substitute the material on one side of the '$\iff$' for the material on the other side, even in a part of a formula.)

b) Use Distribution and Commutation to move all disjunction signs outside conjunctions (for disjunctive normal form) or inside conjunctions (for conjunctive normal form).

Distribution

$$\phi \vee (\psi \wedge \xi) \iff (\phi \vee \psi) \wedge (\phi \vee \xi)$$
$$\phi \wedge (\psi \vee \xi) \iff (\phi \wedge \psi) \vee (\phi \wedge \xi)$$

Commutation

$$\phi \vee \psi \iff \psi \vee \phi$$
$$\phi \wedge \psi \iff \psi \wedge \phi$$

c) Use Association to group things to the left. (Remember, $\ulcorner \psi_1 \wedge \psi_2 \wedge \psi_3 \urcorner$ is short for $\ulcorner (\psi_1 \wedge \psi_2) \wedge \psi_3 \urcorner$.)

Association

$$\phi \vee (\psi \vee \xi) \iff (\phi \vee \psi) \vee \xi$$
$$\phi \wedge (\psi \wedge \xi) \iff (\phi \wedge \psi) \wedge \xi$$

Exercise 7.2: Tautological entailments

1. Using De Morgan's laws, Double Negation Elimination, Commutation, Association, and Distribution, put the following sentences into *both* disjunctive normal form and conjunctive normal form. Show your work.

   a) $(P \wedge Q) \vee (P \wedge \neg Q)$

   b) $P \wedge Q \wedge \neg Q$

   c) $\neg(P \vee Q) \wedge \neg(P \vee \neg Q)$

2. Use the procedure described in §7.3.1 to determine whether the following classical entailments are tautological entailments:

   a) $(P \vee Q) \wedge \neg P \Rightarrow Q$

   b) $P \wedge \neg Q \Rightarrow P$

   c) $P \Rightarrow (P \wedge Q) \vee (P \wedge \neg Q)$

   d) $\neg(P \vee \neg Q) \Rightarrow \neg(P \vee (P \wedge R))$

   e) $A \Rightarrow A \wedge (B \vee \neg B)$

   f) $(A \wedge \neg A) \vee B \Rightarrow B$

It is always possible to reduce a formula to disjunctive normal form and conjunctive normal form in this way. And although there is not a unique normal form for each entailment, it can be proved that if one normal form passes the test for tautological entailment, they all do.

Let's try an example:

$$Q \wedge \neg(\neg P \wedge (Q \vee \neg Q)) \Rightarrow \neg Q \vee P.$$

First, we need to put the premise into disjunctive normal form. Start by driving negations inward using De Morgan's laws:

$$Q \wedge (\neg\neg P \vee \neg(Q \vee \neg Q))$$

$$Q \wedge (\neg\neg P \vee (\neg Q \wedge \neg\neg Q))$$

Use Double Negation Elimination:

$$Q \wedge (P \vee (\neg Q \wedge Q))$$

Now use Distribution whenever we have a disjunction inside a conjunction:

$$(Q \wedge P) \vee (Q \wedge (\neg Q \wedge Q)).$$

Finally, use Association to group to the left:

$$(Q \wedge P) \vee ((Q \wedge \neg Q) \wedge Q)$$

which by convention can be written

$$(Q \wedge P) \vee (Q \wedge \neg Q \wedge Q).$$

Next, we need to put the conclusion

$$\neg Q \vee P$$

into conjunctive normal form. But nothing needs to be done here: it is already in conjunctive normal form. Our reduced entailment statement is therefore

$$(Q \wedge P) \vee (Q \wedge \neg Q \wedge Q) \Rightarrow \neg Q \vee P$$

To see if this is a tautological entailment, we have to check to see if each pair of a disjunct from the premise and a conjunct from the conclusion is explicitly tautological (has a shared atom). Since there are two disjuncts on the left and one conjunct on the right, there are two pairs to consider:

Disjunct	Conjunct	Explicitly tautological?
$Q \wedge \boxed{\text{P}}$	$\neg Q \vee \boxed{\text{P}}$	yes
$Q \wedge \boxed{\neg\text{Q}} \wedge Q$	$\boxed{\neg\text{Q}} \vee P$	yes

Both combinations are explicitly tautological, so the whole thing is a tautological entailment.

7.3.2 The four-valued truth tables

It turns out that the logic of tautological entailment can be captured using four-valued truth tables (Anderson, Belnap, and Dunn 1992, §81). The four truth values are sets of regular truth values: $\{T\}$, $\{F\}$, $\{\}$, $\{T, F\}$. Here are the truth tables for $\neg$ and $\wedge$:

$\neg$	$\{\}$	$\{F\}$	$\{T\}$	$\{T, F\}$
	$\{\}$	$\{T\}$	$\{F\}$	$\{T, F\}$

Exercise 7.3: Truth tables for first-degree entailment

1. What should the table for $\vee$ look like? Study the table for $\wedge$ and figure out the principles behind its construction, and apply these to $\vee$. Explain your reasoning.
2. Use your tables, and the definition of tautological entailment above, to test (3a) and (3b) for tautological entailment. You don't need to give the whole truth table (which can be pretty large with a four-valued logic), but be sure to show your work.

$\wedge$	$\{\}$	$\{F\}$	$\{T\}$	$\{T, F\}$
$\{\}$	$\{\}$	$\{F\}$	$\{\}$	$\{F\}$
$\{F\}$	$\{F\}$	$\{F\}$	$\{F\}$	$\{F\}$
$\{T\}$	$\{\}$	$\{F\}$	$\{T\}$	$\{T, F\}$
$\{T, F\}$	$\{F\}$	$\{F\}$	$\{T, F\}$	$\{T, F\}$

$\phi \Longrightarrow \psi$ is a tautological entailment iff any assignment of values to propositional constants that makes ϕ *at least* T makes ψ *at least* T, and any assignment of values to propositional constants that makes ψ *at least* F makes ϕ *at least* F. (The values $\{T\}$ and $\{T, F\}$ are at least T, and $\{F\}$ and $\{T, F\}$ are at least F.) In other words: validity is preservation of truth and non-falsity.

7.4 Logic and reasoning

Recommended reading

Gilbert Harman, "Logic and Reasoning" (Harman 1984), §1.

Some relevance logicians motivate their view by arguing that classical logic is a bad theory of correct reasoning. Witness Graham Priest:

> the notion of validity that comes out of the orthodox account is a strangely perverse one according to which any rule whose conclusion is a logical truth is valid and, conversely, any rule whose premises contain a contradiction is valid. By a process that does not fall far short of indoctrination most logicians have now had their sensibilities dulled to these glaring anomalies. However, this is possible only because logicians have also forgotten that

> logic is a normative subject: it is supposed to provide an account of correct reasoning. When seen in this light the full force of these absurdities can be appreciated. Anyone who actually reasoned from an arbitrary premise to, e.g., the infinity of prime numbers, would not last long in an undergraduate mathematics course. (Priest 1979, p. 297)

Let's grant that no mathematician would react to the discovery that she had accepted inconsistent premises by concluding (say) that there are infinitely many twin primes.[3] But what does that show?

Harman thinks that the whole line of thought we can find in Priest is based on an overly simplistic assumption about the relation between logic and reasoning.

We can distinguish between two things that might be meant by 'inference':

Reasoning Inference in the broad sense is "reasoned change in view:" revision of beliefs in light of new information or reflection.

Argument Inference in the narrow sense is a process of drawing out the consequences of a given set of premises, in isolation from one's other beliefs.

Reasoning is belief revision. It can involve both additions and subtractions to one's body of beliefs. Argument is what is modeled by practices of formal proof. It is monotonic: adding premises cannot spoil conclusions we already have. It is a fundamental mistake, Harman argues, to confuse the artificial practice of Argument with the natural practice of Reasoning, or belief revision. If you confuse these, you'll think that it's obvious that logic has a special role to play in reasoning. But we shouldn't confuse them. Reasoning is not the same as argument or proof.

For example, it's a confusion to think that Modus Ponens is a rule of Reasoning. Once on a train I heard a young boy exclaim, with some alarm,

(5) I have no pulse!

The boy may have believed

(6) If I have no pulse, I am dead.

But he would obviously not be reasoning well if he applied Modus Ponens to (6) and (5) and came to believe

(7) I am dead.

Dead people cannot check their own pulses. He should instead reexamine his beliefs (6) and (5), and give up one (or both) of them.

Modus Ponens, then, is not a norm for correct Reasoning. It is a rule of Argument. In the formal process of Argument—where one is simply drawing out

[3] Twin primes are pairs of prime numbers that differ by 2, such as 5 and 7.

consequences of some premises, not revising one's views—it is fine to write down (7) when you've written down (6) and (5).

Priest claims that nobody would (or should) reason from a contradiction to, say, the conclusion that there are infinitely many twin primes. This is plausible if we are talking about Reasoning. On discovering a contradiction in our beliefs, we should give up one or both of the contradictory beliefs rather than accepting an arbitrary conclusion. But it doesn't follow from the fact that Ex Falso Quodlibet is a bad rule of Reasoning that it is a bad rule of Argument. (As Harman notes, Modus Ponens is also a bad rule of Reasoning, and the same could be said about any other logical principle the relevantists accept.)

Instead of appealing to intuitions about Reasoning in order to draw conclusions about Argument, could Priest appeal directly to intuitions about Argument? It is not so obvious that people *have* intuitions about Argument that are not just a reflection of their logical training. In an introductory logic course, it takes quite a bit of time to get across the idea that we're trying to capture a kind of "validity" that is independent of the truth of the premises or the plausibility of the conclusion. Think about what needs to be learned when you learn how to "draw out consequences" in this refined sense: you need to learn to ignore the obvious implausibility of the steps you're generating, to ignore things you know to be true that aren't among the premises, to ignore the falsity of the premises. Try asking untrained people whether 'the moon is made of green cheese' can be inferred from 'everything in the sky is made of green cheese' and 'the moon is in the sky', and see what range of responses you get. Of course, people who have been trained in a formal, artificial practice of inferring will immediately answer yes. But they will also say that anything can be inferred from a contradiction—assuming they were trained in classical logic.

By contrast, people *do* have intuitions, independently of their logical training, about how it is correct to infer in the broad sense. But it is not clear how to move from normative claims about belief revision to claims about logical entailment. We would need a bridge principle connecting the two domains. Two obvious ones are

> Ought-believe If you believe A and believe B, and A and B together entail C, you ought to believe C.

> Entitled-believe If you believe A and believe B, and A and B together entail C, you are entitled to believe C.

Clearly these would help to rule out Ex Falso Quodlibet. Indeed, Priest seems to be appealing implicitly to something like these principles. But as Harman argues, neither is plausible.

Ought-believe requires you to believe many trivialities, and sets an impossible standard. To conform to this norm, you're required to believe every true mathematical theorem if you believe the axioms.[4]

Entitled-believe doesn't *require* you to believe all these things; it just permits it. But it doesn't take into account the possibility that, on recognizing that you believe *A* and *B* which together entail *C*, you might sometimes be required to reject *C* and stop believing either *A* or *B* or both. (See the argument from (6) and (5) to (7), above.)

One might fall back to the view that logic gives us no positive guidance, but only *negative* guides to belief revision:

Ought-not-believe-strong If *A* and *B* together entail *C*, you ought not believe all of *A*, *B*, and not-*C*.

Ought-not-believe-weak If you believe that *A* and *B* together entail *C*, you ought not believe all of *A*, *B*, and not-*C*.

These principles allow that we may choose between rejecting a premise and accepting the conclusion of a valid argument, but they require us to do one of these things.

However, Harman criticizes even these "negative" norms:

> On discovering one has inconsistent beliefs, one might not see any easy way to modify one's beliefs so as to avoid the inconsistency, and one may not have the time or ability to figure out the best response. In that case, one should (at least sometimes) simply acquiesce in the contradiction while trying to keep it fairly isolated. I would think this is the proper attitude for most ordinary people to take toward many paradoxical arguments.
>
> Furthermore, a rational fallible person ought to believe that at least one of his or her beliefs is false. But then not all of his or her beliefs can be true, since, if all of the other beliefs are true, this last one will be false. So in this sense a rational person's beliefs are inconsistent. (Harman 1984, pp. 108–9)

When you find that you have an inconsistent set of beliefs but can't identify the source of the inconsistency, does rationality really require that you give up all the beliefs, or all but one? These are the only revisions you can *know* will restore consistency.

Perhaps there is some connection between logical implication/entailment and belief revision, but it is not at all obvious. But if we can't argue from robust intuitions about belief revision to the rejection of Ex Falso Quodlibet, how can we argue for it?

[4] Could this be fixed by inserting "you believe that" before "*A* and *B* together entail *C*"?

7.5 Uses for relevance logic

The initial arguments for relevance logic were willing to grant that classical forms of inference preserved truth. The idea was that something more was required, in addition to truth preservation: the premises had to be relevant to the conclusion.

But this position is hard to maintain. If we concede that disjunctive syllogism preserves truth, then we can derive disjunctive syllogism, given Modus Ponens and some plausible assumptions about the truth predicate:

1	$(\mathrm{Tr}(\text{`}P \vee Q\text{'}) \wedge \mathrm{Tr}(\text{`}\neg P\text{'})) \rightarrow \mathrm{Tr}(\text{`}Q\text{'})$	premise
2	$P \vee Q \rightarrow \mathrm{Tr}(\text{`}P \vee Q\text{'})$	premise
3	$\neg P \rightarrow \mathrm{Tr}(\text{`}\neg P\text{'})$	premise
4	$\mathrm{Tr}(\text{`}Q\text{'}) \rightarrow Q$	premise
5	$P \vee Q$	Hyp
6	$\neg P$	Hyp
7	$\mathrm{Tr}(\text{`}P \vee Q\text{'})$	Modus Ponens 2, 5
8	$\mathrm{Tr}(\text{`}\neg P\text{'})$	Modus Ponens 3, 6
9	$\mathrm{Tr}(\text{`}P \vee Q\text{'}) \wedge \mathrm{Tr}(\text{`}\neg P\text{'})$	$\wedge$ Intro 7, 8
10	$\mathrm{Tr}(\text{`}Q\text{'})$	Modus Ponens 1, 9
11	Q	Modus Ponens 4, 10

(7.1)

Premise 1 is the claim that Disjunctive Syllogism preserves truth; premises 2, 3, and 4 follow from common assumptions about the truth predicate. The argument uses only Modus Ponens and $\wedge$ Intro, both of which are relevantly valid.

More recently, relevantists have not presented relevance as an additional condition for validity, on top of truth preservation. They have argued, instead, that relevance logics are to be preferred because classical inferences forms do not preserve truth. The idea is to keep the classical idea that validity is truth preservation, but give up the classical assumption that the same sentence cannot be both true and false.

This approach is suggested by the four-valued truth tables. In Ex Falso Quodlibet, you can go from a premise that is $\{T, F\}$ to a conclusion that is $\{F\}$. If we think of $\{T, F\}$ as "both true and false" and $\{F\}$ as "just false," we have gone from a premise that is true to one that is not. So truth is not preserved.

But does it make sense to think of $\{T, F\}$ as "both true and false"? We need to think more about what these values mean.

7.5.1 Dialetheism

Recommended reading

Graham Priest, "What Is So Bad About Contradictions?" (Priest 1998).

Dialetheism is the view that some propositions are both true and false (that is, both they and their negations are true). This is decidedly a minority view, but Priest (1998) defends it. On Priest's view, Ex Falso Quodlibet is to be rejected because it fails to preserve truth: '$P \wedge \neg P$' can be both true and false when 'Q' is just plain false.[5]

To support his view, Priest gives some examples of sentences that are both true and false. The first is the classic Liar paradox:

(8) This sentence is not true.

Using some simple principles governing the truth predicate, we can prove that if (8) is true, it is false, and if it is false, it is true. Paradox ensues under the assumption that no sentence can be both true and false, but the paradox vanishes if we allow (8) to have both truth values.

A second example assumes that the department's decision makes it the case that a dissertation is accepted (or not). Suppose that through some mistake, the authorities say both that Sarah's dissertation is accepted and that it is not. Then, Priest argues,

(9) Sarah's dissertation is accepted

is both true and false.

A final example. Suppose my center of gravity is on the vertical plane containing the center of gravity of the door. By the symmetry of the situation, either I'm both in and out of the room, or neither in nor out. But if I'm neither in nor out, then it follows that I'm both in and out (assuming that if I'm in, I'm not out and vice versa, and applying De Morgan's laws and Double Negation Elimination). So

(10) I am in the room.

[5] Priest advocates a logic LP that is *paraconsistent*—meaning that it fails to validate Ex Falso Quodlibet—but not *relevant*, since it validates the inference from '$P \wedge \neg P$' to '$Q \vee \neg Q$'. All relevance logics are paraconsistent, but not all paraconsistent logics are relevance logics.

is both true and false.

It is surprisingly difficult to find a non-question-begging argument against dialetheism. David Lewis has famously argued that we should decline to try:

> The reason we should reject this proposal is simple. No truth does have, and no truth could have, a true negation. Nothing is, and nothing could be, literally both true and false. This we know for certain, and *a priori*, and without any exception for especially perplexing subject matters. The radical case for relevance should be dismissed just because the hypothesis it requires us to entertain is inconsistent.
>
> That may seem dogmatic. And it is: I am affirming the very thesis that Routley and Priest have called into question and—contrary to the rules of debate—I decline to defend it. Further, I concede that it is indefensible against their challenge. They have called so much into question that I have no foothold on undisputed ground. So much the worse for the demand that philosophers always must be ready to defend their theses under the rules of debate. (Lewis 1998, p. 101)

7.5.2 The moderate approach

Recommended reading

Alan Ross Anderson, Nuel D. Belnap, Jr., and J. Michael Dunn, "A Useful Four-Valued Logic," (Anderson, Belnap, and Dunn 1992, §§81.1–2).

Anderson, Belnap, and Dunn (1992) give a somewhat different kind of argument. They are not here arguing for full-blown relevantism: the view that we should use relevance logic for all purposes. Rather, they argue that relevance logic is preferable to classical logic for extracting information from a database that might contain inconsistent information. Imagine that the database is fed information from several fairly reliable (but not infallible) sources. We want to be able to ask it questions and get useful answers.

Let's say the sources *only* report on atomic sentences. Each source says either T or F or nothing about a given sentence. So with respect to each sentence the computer can be in four possible states:

$\{T\}$	told true only	$\{F\}$	told false only
$\{T, F\}$	told both true and false	$\{\}$	told neither

We want to be able to ask the computer not just about atomic sentences (and here it can just spit out the value) but about compounds. To answer our questions, it needs to do some *deduction*. Anderson, Belnap and Dunn propose the four-valued logic $\mathsf{E}_{\mathsf{fde}}$ as a procedure the computer can be programmed to follow in doing this "reasoning."

Why not just use classical logic? Remember that the database may contain inconsistent information. For a given atomic sentence, it might have been told both true and false, so it might reckon both the sentence and its negation true. If it uses classical logic as a canon for reasoning, then as soon as it gets into an inconsistent state, it will start saying "Yes" to every question. That would make the database useless.

Anderson, Belnap and Dunn concede that it would be preferable to program the computer so that it could subtract beliefs as well as adding them, as in human belief revision, but in the absence of a complete algorithm for belief revision, a relevance logic is a good second best:

> The complete reasoner should, presumably, have some strategy for *giving up* part of what it believes when it finds its beliefs inconsistent. Since we have never heard of a practical, reasonable, mechanizable strategy for revision of belief in the presence of contradiction, we can hardly be faulted for not providing our computer with such. In the meantime, while others work on this extremely important problem, our computer can only accept and report contradictions without divesting itself of them. (Anderson, Belnap, and Dunn 1992, p. 508)

We can see why, in this context, it makes sense to reject Disjunctive Syllogism. Suppose the database has been told both that 'P' is true and that 'P' is false, but has been told only that 'Q' is false. Then '$P \vee Q$' will be $\{T, F\}$, 'P' will be $\{T, F\}$, and 'Q' will be $\{F\}$. Using Disjunctive Syllogism would therefore allow one to go from premises that are told-true to a conclusion that is not told-true.

However, the inference from '$(A \vee B) \wedge \neg A$' to '$(A \wedge \neg A) \vee B$' is valid in $\mathsf{E}_{\mathsf{fde}}$.

> That is, having determined that the antecedent is at least told True, we allow the computer to conclude: either B is at least told True, or something funny is going on; i.e., it's been told that A is both True and False. And this, you will see, is right on target. If the *reason* that $(A \vee B) \wedge \neg A$ is getting thought of as a Truth is because A has been labeled as both told True and told False, then we certainly do *not* want to go around inferring B. The inference is wholly inappropriate in a context where inconsistency is a live possibility. (Anderson, Belnap, and Dunn 1992, p. 520, with minor notational adjustments)

7.5.3 Truth in a corpus

Recommended reading

David Lewis, "Logic for Equivocators" (Lewis 1998).

Lewis (1998) likes the idea of trying to make sense of "truth in a (possibly inconsistent) corpus," but thinks relevance logic isn't the right way to do it. He prefers something he calls "compartmentalization."

Lewis proposes the following desiderata for a notion of "truth in a corpus":

Truth in a corpus

1. Anything explicitly affirmed is true in the corpus.
2. There is more in the corpus besides what is explicitly affirmed—truth in a corpus is to some extent closed under implication.
3. But a corpus can contain inconsistency without containing everything.
4. ϕ is false in the corpus iff $\ulcorner\neg\phi\urcorner$ is true in the corpus.
5. The orthodox rules for '$\wedge$' and '$\vee$' apply without exception.

Lewis agrees that if you want 1–5, you get something like the logic E. But he thinks a natural, useful conception of truth in a corpus will reject 5.

> I am inclined to think that when we are forced to tolerate inconsistencies in our beliefs, theories, stories, etc., we quarantine the inconsistencies entirely by fragmentation and not at all by restrictions of relevance. In other words, truth according to any single fragment is closed under unrestricted classical implication.
>
> [E] cannot be trusted to preserve truth according to a fragmented corpus, nor can any logic that ever lets us mix fragments in many-premise implications. (Lewis 1998, p. 105)

He illustrates this with the following example. At one time he believed:

(11) Nassau Street runs roughly East–West.

(12) The railroad runs roughly North–South.

(13) Nassau Street is roughly parallel to the railroad.

But, he says, he didn't believe their conjunction—that Nassua Street ran roughly East–West *and* the railroad ran roughly North–South *and* they were roughly parallel. This wasn't true according to the "corpus" of his beliefs. But it would be if truth-on-a-corpus were closed under entailment in E: relevantists have no beef with conjunction introduction.

Instead of thinking of the corpus as a unified whole, closed under some nonclassical entailment relation, Lewis thinks of it as broken up or "compartmentalized" into overlapping fragments. Different pieces are active in reasoning in different situations, but never all of these at once. Triviality is avoided by allowing inferences only when all the premises are in the same fragment. Within a fragment, all classical inferences are valid (even Disjunctive Syllogism).

Having argued that relevance logic is not needed to deal with inconsistent corpora, Lewis suggests that if it is good for anything, it is for preventing damage from potential equivocation:

> We teach logic students to beware of fallacies of equivocation. It would not do, for instance, to accept the premise $A \vee B$ because it is true on another disambiguation of A, and then draw the conclusion B. After all, B might be unambiguously false. The recommended remedy is to make sure that everything is fully disambiguated before one applies the methods of logic.
>
> The pessimist might well complain that this remedy is a counsel of perfection, unattainable in practice. He might say: ambiguity does not stop with a few scattered pairs of unrelated homonyms. It includes all sorts of semantic indeterminacy, open texture, vagueness, and whatnot, and these pervade all of our language. Ambiguity is everywhere. There is no unambiguous language for us to use in disambiguating the ambiguous language. So never, or hardly ever, do we disambiguate anything fully. So we cannot escape fallacies of equivocation by disambiguating everything. Let us rather escape them by weakening our logic so that it tolerates ambiguity; and this we can do, it turns out, by adopting some of the strictures of the relevantists. (Lewis 1998, pp. 107–8)

Read *true osd* as "true on some disambiguation." Then we get three values: *true-osd only*, *false-osd only*, *both true-osd and false-osd*. If we define validity as preservation of truth-osd, we get Priest's LP; if we define it as preservation of truth-osd and non falsity-osd, we get the relevance logic R-mingle.

Relevance logic, on this view, is "logic for equivocators." It is the logic we should use if we fear that we have been using our words—'red', 'terrorist', 'qualified'—in subtly different senses at different points in our argumentation.

Further readings

- The classic here is the two-volume *Entailment* (Anderson and Belnap 1975; Anderson, Belnap, and Dunn 1992), which examines relevance logic in great detail and from many angles.
- Dunn and Restall 2002 is a more recent survey article.
- Read 1988 is also useful because of its attention to the differences between the American, Australian, and Scottish schools of relevance logic.
- On the issues raised by Harman about logic and norms for reasoning, see Foley 1992, Christensen 2004, Field 2009b, and Steinberger 2017.

8 Vagueness and the Sorites Paradox

In the semantics for classical logic, the interpretation of a predicate is a set of objects—the objects to which the predicate applies. This approach presupposes that, for each object in the domain, there is a fact of the matter as to whether the predicate applies to it. Is this presupposition tenable for vague predicates like 'tall' or 'stale'? If so, there must be a shortest tall building (or several that are tied), and a particular microsecond at which a doughnut becomes stale. But it seems hard to believe that words like 'tall' and 'stale' make such fine distinctions. Many philosophical logicians have concluded that classical logic applies only in the domain of the precise: mathematics and science.[1]

It is natural to ask whether relaxing some of the assumptions of classical logic and semantics, such as the assumption that every sentence is either true or false, could yield a logic suitable for analyzing arguments that use vague language. In this chapter we will consider several influential proposals along these lines: three-valued logic, fuzzy logic, and supervaluational semantics. We will conclude by considering whether vagueness is a feature of the world itself or merely of our ways of describing it, looking at a famous argument of Gareth Evans that bears on this question.

8.1 What is vagueness?

There is probably no way of defining vagueness that will be acceptable to all theorists. But three features are standardly associated with vagueness: borderline cases, a lack of sharp boundaries, and susceptibility to sorites arguments.

Borderline cases

In standard semantics, we assume that each predicate partitions the domain into two sets: those to which the predicate applies, and those to which it does not. But

[1] One proponent of this view was Gottlob Frege, the father of modern logic. For an illuminating discussion of Frege's views, see Puryear 2013.

for vague predicates, there seem to be borderline cases. Is a pile of 20 grains of sand a heap? Is someone with 100 hairs on his head bald?

Epistemicists like Williamson (1994) have argued that even in borderline cases, there is a fact of the matter as to whether the predicate applies. If they are right, then there is no difficulty applying classical semantics and logic to vague discourse. The epistemicists' burden is to explain how it is that our vague words acquire these classical extensions, why we cannot know where the boundaries lie, and how we can communicate with words whose meanings are, in an important sense, beyond our ken. These are problems in epistemology and the philosophy of language. But, because epistemicism requires no innovations in logic, we will not discuss the position further in this book.

Anti-epistemicists think that our hesitation to answer questions about whether a pile of 20 grains of sand a heap, or whether a man with 100 hairs on his head is bald, is not just a matter of ignorance. Rather, they think, there is "no fact of the matter" about whether these borderline cases fall into the extension of the vague term. (But what, exactly, does that mean?)

No sharp boundaries

In principle there might be a word with borderline cases, but sharp boundaries between them and the clear cases. Imagine an artificial predicate, 'yuve', defined as follows:

> For all x, x is yuve if x is less than 16 years old, and x is not yuve if x is greater than 21 years old.

One might consider an 18-year old a borderline case of a yuve person, as nothing determines whether an 18-year old is yuve. But 'yuve' does not seem to be vague. For vagueness, we need not just borderline cases, but "no sharp boundaries" between the definite cases and the borderline cases.

The sorites

Vague terms are susceptible to *sorites arguments*. The term 'sorites' means *heaper* in ancient Greek. The original form of the paradox assumed that removing a single grain of sand from a heap still leaves you with a heap, and concluded that a single grain of sand forms a heap. We can give a more contemporary version by defining the predicate R as follows:

$$R(x) \equiv_{def} x \text{ cents is enough to make you rich.}$$

Now consider the argument:

$$(1)\quad \begin{array}{l} \forall n(R(n) \rightarrow R(n-1)) \\ R(1{,}000{,}000{,}000) \\ \hline R(1) \end{array}$$

The argument has two premises, both of which seem hard to deny. But a series of logically valid steps (∀ Elim and Modus Ponens) brings us to a conclusion that seems manifestly false. A paradox!

Instead of using a universally quantified premise, we can give a version of the argument that uses 999,999,999 conditional premises:

$$(2)\quad \begin{array}{l} R(1{,}000{,}000{,}000) \rightarrow R(999{,}999{,}999) \\ R(999{,}999{,}999) \rightarrow R(999{,}999{,}998) \\ \ldots \\ R(2) \rightarrow R(1) \\ R(1{,}000{,}000{,}000) \\ \hline R(1) \end{array}$$

All of these premises seem hard to deny. Here, the only rule of inference needed to get the conclusion is Modus Ponens.

These arguments are challenging. The premises and argumentation are compelling, but we cannot accept the conclusion. If we don't want to accept the conclusion, we must either reject one of the premises or reject one of the modes of inference used. Rejecting the premise that a billion cents is enough to make you rich seems just as absurd as rejecting the conclusion. What about the other premises?

In classical logic, rejecting the truth of the first premise of (1) means accepting the truth of its negation,

$$\neg\forall n(R(n) \rightarrow R(n-1))$$

which is classically equivalent to

$$\exists n(R(n) \wedge \neg R(n-1)).$$

This posits a sharp one-penny boundary between the rich and the non-rich, and that is difficult to swallow.

Similarly, if we want to reject (2), we need to reject at least one conditional of the form

$$R(n) \rightarrow R(n-1).$$

In classical logic, that means accepting the conjunction

$$R(n) \wedge \neg R(n-1)$$

for some particular n. And again, we are forced to accept that there is a sharp one-penny boundary between the rich and the non-rich.

Epistemicists bite these bullets, accepting that there is a particular n such that n pennies makes you rich but $n - 1$ pennies does not. If we want to avoid this conclusion, we need to ask whether giving up some principles of classical logic and semantics will allow us to do so.

8.2 Three-valued logics

Recommended reading

Timothy Williamson, *Vagueness*, §§4.1–4.6 (Williamson 1994).

An ideal solution to the sorites paradox would allow us to reject the problematic premises as untrue, without accepting their negations as true. But that isn't possible in classical logic, because it is committed to

Bivalence Every sentence is either true or false.

Given bivalence, if the premise is not true, it is false, and its negation is true.

But suppose we drop bivalence and allow that, in the borderline area, the conditionals are neither true nor false. The most obvious way to do this is to introduce a third truth value: in addition to T (true) and F (false), we introduce a value N (neither). This gives us a *three-valued logic*.

8.2.1 Semantics for connectives

A model is an assignment of one of these values (T, F, N) to each propositional constant. The values of compound formulas are determined, as in classical propositional logic, by truth tables.

What should our three-valued truth tables look like? Here there is some controversy. Everyone agrees with the following constraints (considering just negation and conjunction):

1. The tables should agree with classical tables on classical "inputs" (T or F)
2. Negation should take T to F, F to T, and N to N.
3. $A \wedge A$ should have the same value as A.

$\wedge$	T	N	F
T	T	N	F
N	N	N	N
F	F	N	F

$\vee$	T	N	F
T	T	N	F
N	N	N	N
F	F	N	F

$\neg$	
T	F
N	N
F	T

Figure 8.1: Weak Kleene tables.

$\wedge$	T	N	F
T	T	N	F
N	N	N	F
F	F	F	F

$\vee$	T	N	F
T	T	T	F
N	T	N	N
F	F	N	F

$\neg$	
T	F
N	N
F	T

Figure 8.2: Strong Kleene tables.

4. $A \wedge B$ should have the same value as $B \wedge A$.

These constraints leave only four squares where there is room for disagreement:

$\wedge$	T	N	F
T	T	?	F
N	?	N	?
F	F	?	F

There are two main strategies for filling in the squares marked with '?'. One is to treat N as "infectious," so that if any conjunct is N, the conjunction is N. (This makes sense if you read N as "meaningless.") The resulting truth tables are usually called the *Weak Kleene* tables, after Kleene (1952, §64) (see Fig. 8.1).

If, instead, we think of N as "undetermined," a different strategy is appropriate. The truth of one disjunct is sufficient for the truth of a disjunction, so $T \vee N$ should be T, not N. For similar reasons, $F \wedge N$ should be F. When we use this reasoning to fill in the squares, we get the *Strong Kleene* tables (see Fig. 8.2).

8.2.2 Defining validity in multivalued logics

We have truth tables, but still no logic, because we have not said what it is for an argument to be valid in these logics. Here we face an interesting choice. In classical logic, there is no difference between three definitions of validity:

a) *preservation of truth*
$\Gamma \models_a B$ iff for every model v, if every formula in Γ is T on v, then B is T on v.

b) *preservation of non-falsity*
$\Gamma \models_b B$ iff for every model v, if every formula in Γ is non-F (T or N) on v, then B is non-F on v.

c) *preservation truth and non-falsity*
$\Gamma \models_c B$ iff $\Gamma \models_a B$ and $\Gamma \models_b B$.

All of these come to the same when we have bivalence. But when we introduce a third truth value, they come apart, and we have to choose. Our choice may be based both on philosophical considerations—what do we really care about preserving when we use valid arguments?—and on consideration of particular cases.

Sometimes validity in multivalued logics is defined as the preservation of *designatedness*. Certain values are "designated," and an inference form is regarded as valid if no instance can go from premises with designated values to a conclusion with a non-designated value. In a three-valued logic, designating T yields Γ_a, and designating T and N yields Γ_b.[2]

8.2.3 Application to the sorites

How could a three-valued logic help with the sorites paradox? Assume the Strong Kleene truth tables. Since we haven't presented a semantics for the universal quantifier, we will consider the form of the sorites that just uses conditionals (2, above). We'll understand the conditionals as material conditionals, defined in the usual way:

$$(A \rightarrow B) \equiv_{\text{def}} (\neg A \vee B).$$

The conditionals at the beginning, like

$$R(1{,}000{,}000{,}000) \rightarrow R(999{,}999{,}999)$$

[2] Of the authors discussed by Williamson, Łukasiewicz, Bochvar, Kleene, and Tye opt for preservation of truth (designating only T), while Halldén opts for preservation of non-falsity (designating T and N).

Exercise 8.1: Three-valued logics

1. For each of these inferences

   i) $\dfrac{A \vee B \quad \neg A}{B}$ ii) $\dfrac{A}{\neg B \vee (A \wedge B)}$

   determine whether it is valid on the Strong Kleene tables, on each of the three definitions of validity: (a) preservation of truth, (b) preservation of non-falsity, and (c) preservation of truth and non-falsity.

2. Do the same thing with the Weak Kleene tables.

3. Are there any interesting logical relations between our three notions of validity, with the Strong Kleene tables? Clearly

   $$\Gamma \models_c B \Rightarrow \Gamma \models_a B$$
   $$\Gamma \models_c B \Rightarrow \Gamma \models_b B.$$

   But are there any other entailments that hold?

4. The Weak and Strong Kleene tables, together with any of our three definitions of validity, yield a nonclassical logic. Can you come up with a three-valued semantics that yields classical propositional logic?

will have T antecedents and T consequents, and will therefore be T. The conditionals at the end, like

$$R(3) \rightarrow R(2)$$

will have F antecedents and F consequents, and will therefore be F. But in the middle, at some point, we'll have a conditional with a T antecedent and an N consequent (which will be N), followed by a number of conditionals with N antecedents and N consequents (which will also be N), and a conditional with an N antecedent and a F consequent (which will be N too). So, most of our premises will be T, and none will be F, but somewhere in the middle we'll have a number of premises that are N.

We now have a choice, depending on how we define validity:

- If we define validity as the preservation of truth, then Modus Ponens is valid, so the argument in (2) is valid. We end up with a false conclusion because, on this understanding, valid arguments need not preserve non-falsity. If we had only T premises, the conclusion could not be F, but with some N premises, anything is possible.

- On the other hand, if we define validity as the preservation of truth and non-falsity, or as the preservation of non-falsity, then Modus Ponens is not valid. (When p is N and q is F, $\ulcorner p \rightarrow q \urcorner$ will be N.) So, on this way of looking at things, a valid argument wouldn't let you go from premises that are all T or N to a conclusion that is F, but the sorites argument is not valid.

Either way, we have provided a way to reject some of the premises of the sorites argument without accepting a "sharp boundaries" claim of the form: "n cents is enough to make you rich, but $n - 1$ cents is not."

We do, however, still have "sharp boundaries" between the Ts and Ns, and between the Ns and Fs. And they may seem to pose problems of the same kind as before. The problem is especially vivid if we introduce into the language a one-place operator D ("definitely") that would allow us to say that someone who is a borderline case of being rich is "not definitely rich and not definitely not rich."

p	Dp
T	T
N	F
F	F

For we can then reproduce the sorites argument at a higher level, using premises like

$$DR(100{,}000{,}000) \rightarrow DR(99{,}999{,}999)$$

The idea that one cent can't make a difference as to whether you have enough money to be *definitely rich* seems as plausible as the idea that it can't make a difference to whether you have enough money to be rich. So are we really better off with a three-valued logic than a two-valued one?

8.3 Fuzzy logics

Recommended reading

Timothy Williamson, *Vagueness*, §§4.7–4.8, 4.10, 4.14 (Williamson 1994).

We might hope to solve the problem by allowing that truth can come in degrees. As we get to the borderline area, the conditionals become a bit less than fully true. Because Modus Ponens does not quite preserve degree of truth, you can go from lots of *almost* fully true premises to a false conclusion. A semantics based on the idea that truth comes in degrees yields what is sometimes called "fuzzy logic."

8.3.1 Semantics

On this approach, the truth values are real numbers between 0 and 1 (inclusive). A model is an assignment of these values to propositional constants. These assignments can be extended to compound formulas using the rules below.[3] Here $|\phi|$ is the degree of truth of ϕ, max is the function that returns the greater of two numbers (or the first, if they are equal), and min is the function that returns the lesser of two numbers.

$$
\begin{aligned}
|\neg A| &= 1 - |A| \\
|A \vee B| &= \max(|A|, |B|) \\
|A \wedge B| &= \min(|A|, |B|) \\
|A \rightarrow B| &= \begin{cases} 1 & \text{if } |A| \leq |B| \\ 1 - (|A| - |B|) & \text{otherwise} \end{cases} \\
|A \equiv B| &= 1 - (\max(|A|, |B|) - \min(|A|, |B|))
\end{aligned}
$$

When it comes to defining validity, we again face a choice, analogous to the choice we faced with three-valued logics:

a) *preservation of perfect truth*
   $\Gamma \models_p B$ iff for every model v, if every formula in Γ is 1 on v, then B is 1 on v.

b) *preservation of degree of truth*
   $\Gamma \models_d B$ iff for every model v, the value of B on $v \geq$ the lowest value of a member of Γ on v.

8.3.2 Application to the sorites

How does a fuzzy semantics help with the sorites paradox? The premises of the sorites have the form

$$R(n) \rightarrow R(n-1)$$

[3] These are the most common rules, due originally to Łukasiewicz. For some different approaches, see Goguen 1969, Edgington 1997, and Weatherson 2005. Edgington's semantics is not degree-functional.

with different values of n. In each case, the degree of truth of the consequent will be the same as, or only very slightly less than, the degree of truth of the antecedent. In the former case, the conditional will have degree 1; in the latter case, it will have a degree very close to 1. Thus, on this approach, all of the conditionals are either perfectly true or nearly perfectly true.

Is the argument valid? Since we have a version that only uses Modus Ponens, we just need to ask whether Modus Ponens is valid. As in the three-valued case, it depends on how we define validity:

- If validity is preservation of perfect truth (=1), then Modus Ponens is valid. But, since there is no guarantee that *nearly* perfect truth will be preserved, we can go from a bunch of premises that are either perfectly true or nearly so to a perfectly false conclusion.

- If validity is preservation of degree, then Modus Ponens is not valid. The degree of its conclusion can be less than the degrees of any of the premises. Example: consider the inference from A and $A \rightarrow B$ to B, when A has degree 0.8 and B has degree 0.6. In this case $A \rightarrow B$ will have degree 0.8, so we'll go from two premises with degree 0.8 to a conclusion with degree 0.6.

Either way, we have an explanation of where the argument goes wrong that does not require us to embrace a "sharp boundaries" claim.

However, as Weatherson (2005, p. 62) observes, the explanation falls apart when we reformulate the sorites using negated conjunctions instead of conditionals. Premises of the form

$$\neg(R(n) \wedge \neg R(n-1))$$

seem just as compelling as the conditionals in our original argument. But these premises will vary in truth from 1 (in the clear cases) to 0.5 (in the borderline cases). In this reformulated argument, we can no longer say that all of the premises are either perfectly or nearly perfectly true. Indeed, some of the premises are no more true than false. Why, then, do *all* of them seem compelling? We will need some explanation of their plausibility that doesn't assume that they are true or nearly true. But if our explanation of the plausibility of the premises of the sorites doesn't depend on truth coming in degrees, what are we gaining, exactly, by embracing a fuzzy logic?

8.3.3 Can we make sense of degrees of truth?

Does it make sense to think of truth as coming in degrees? It is common to say that one description is "more true" than another, but this might just mean

that it contains more true propositions. Does it make sense to think of a single proposition as more or less true?

Some degree theorists attempt to explain degrees of truth via comparatives. The general idea is to move from

(3) *X* is balder than *Y*

to

(4) '*X* is bald' is truer than '*Y* is bald'.

Forbes (1985) argues more or less as follows (see Keefe 2000, p. 92):

1. *x* is balder than *y*.
2. *x* is bald to a greater degree than *y*.
3. *x* satisfies 'is bald' to a greater degree than *y* does.
4. '*x* is bald' is truer than '*y* is bald'.

The step from 1 to 2 is fine, and the step from 3 to 4 is underwritten by the conceptual connection between truth and satisfaction. But the step from 2 to 3 can be questioned.

One problem is that this form of argument yields quite counterintuitive consequences for *relative gradable adjectives*. 'Bald' is an *absolute gradable adjective*, which means that it makes sense to say that someone is "completely bald." One can get balder and balder up to a point, but no one can be balder than a person with no hair at all. 'Tall', by contrast, has no such limit. There is no such thing as being "completely tall" or "perfectly tall." It is always possible, in principle, to be taller.

Now consider two clearly tall people: say, Michael Jordan and Bill Clinton. Since Jordan is tall to a greater degree than Clinton, this argument requires us to assign a greater degree of truth to 'Jordan is tall' than to 'Clinton is tall'. This means that 'Clinton is tall' cannot have degree 1, even though Clinton seems a clear case of a tall man. Indeed, 'Jordan is tall' cannot have degree 1, for the same reason. So Forbes' argument, applied to relative gradable adjectives, yields the uncomfortable result that there are no clear cases of tallness.

A further problem is that comparatives can only take us so far. If every sentence is to be assigned a degree of truth, then we must settle whether 'Paris is beautiful' is truer than 'London is big'. These sentences use different gradable adjectives, so we cannot fall back on comparatives to understand what it means for one to be truer than the other.

Can we make sense of degrees of truth without going through comparatives? Mark Sainsbury appeals to our understanding of truth as the aim of belief:

> Truth is what we seek in belief. It is that than which we cannot do better. So where partial confidence is the best that is even theoretically available, we need a corresponding concept of partial truth or degree of truth. Where vagueness is at issue, we must aim at a degree of belief that matches the degree of truth, just as, where there is no vagueness, we must aim to believe just what is true. (Sainsbury 1995, p. 44)

That is: to say that a proposition is true is to say that we would meet our aim in cognition if we believed it fully. To say that a proposition is true to degree d is to say that it we would meet our aim in cognition if we believed it to degree d.

There is something plausible about this idea. In borderline cases, we feel "ambivalence" (Wright 2001, pp. 69–70) about whether a term applies, and one way to think of this ambivalence is as a kind of partial belief or partial endorsement. But is this the same kind of partial belief we have in cases of uncertainty, or something distinct? If it is distinct, how does it relate to uncertainty?[4]

8.3.4 Troubles with degree-functionality

To say that a connective is *degree-functional* is to say that the degree of truth of a compound formula formed with the connective is completely determined by the degrees of truth of the formulas it connects. (This is a generalization of the familiar notion of truth functionality.) Our fuzzy logic is degree-functional. As Williamson notes, this has some counterintuitive consequences.

Suppose Castor and Pollux are about the same height, and that both are borderline cases of being *tall*. Pollux is just a bit taller than Castor—say, by 1 mm. Intuitively, (5a) should be perfectly true while (5b) should be perfectly false (after all, Pollux is taller than Castor):

(5) a. Castor is tall → Pollux is tall
    b. Castor is tall → ¬Pollux is tall.

Similarly, (6a) should be perfectly false, while (6b) should have an intermediate value:

(6) a. Castor is tall ∧ ¬Pollux is tall
    b. Castor is tall ∧ Pollux is tall.

Finally, (7a) should be perfectly true, while (7b) should be perfectly false:

(7) a. Pollux is tall → Pollux is tall
    b. Pollux is tall → ¬Pollux is tall.

[4] For discussion, see Field 2000, Schiffer 2003, ch. 5, MacFarlane 2006, and MacFarlane 2010.

Our fuzzy logic cannot deliver these intuitive results, because it is degree-functional. Suppose that

$$|\text{Pollux is tall}| = 0.5.$$

(That is, Pollux is smack in the middle of the borderline cases.) Then

$$|\text{Pollux is tall}| = |\neg\text{Pollux is tall}|.$$

And that means that we can substitute '¬Pollux is tall' for 'Pollux is tall' in compound sentences without affecting the degree of the compound. Hence the sentences in each pair (5a/5b), (6a/6b), and (7a/7b) must have the same degree of truth, contrary to our intuitive judgments.[5]

In the face of this objection, the fuzzy theorist might push back on the motivating intuitions. If 'Pollux is tall' has degree 0.5, then it is no more true than false, and its negation is no more true than false. If it is no more true than false that Pollux is tall, and no more true than false that he is not tall, why should it be perfectly false that he is tall and not tall? Asked whether Polux is tall, we might naturally respond, "He is and he isn't."

Once we accept that a contradiction like

(8) Pollux is tall ∧ ¬Pollux is tall

can have an intermediate value (0.5), the intuition that (6a) is perfectly false no longer seems so compelling. Surely, if (8) can have an intermediate value, so can (6a). And the concession that 'Pollux is tall' is no more true than false can help dispel the intuition that (7a/7b) and (5a/5b) should differ in degree.

8.4 Supervaluations

Recommended reading

Timothy Williamson, *Vagueness*, §§5.1–5.4 (Williamson 1994).

The supervaluational approach is designed to appeal to those who find the objections to degree functionality discussed in §8.3.4 compelling. Fine (1997), who first advocated this approach, uses the following motivating example. Consider a blob whose color we have trouble classifying either as red or as pink: it is a borderline case. On a degree theory, we might have

$$|\text{Blob is red}| = |\text{Blob is pink}| = 0.5.$$

[5] A similar objection can be made to the three-valued logics we considered in §8.2, above.

But then the degree theorist is committed to

$$|\text{Blob is red} \vee \text{Blob is pink}| = 0.5.$$

This, Fine argues, is wrong. Blob is *clearly* either red or pink: it may be indeterminate which of these shades it has, but it is determinate that it has one of them. So the disjunction

(9) Blob is red or pink

is determinately true. Supervaluationism is designed to deliver this result, allowing a disjunction to be determinately true even though neither disjunct is.

Supervaluationism piggybacks on classical semantics. The idea is that in the presence of vagueness, there is no single "intended" classical model. Rather, there is a range of classical models that respect the vague meanings of our words. In considering whether a sentence is true or false, we should consider all of these models. A sentence is true if it is true on all classical models that respect the partial constraints imposed by the meanings of our words, false if it is false on all such models, and indeterminate otherwise.

A supervaluational *interpretation* of a language, then, is a *set* of classical models, which we will call *valuations*. Recall that a classical model assigns an extension to every nonlogical predicate, relation, and term, and also specifies a domain.

Formally, an interpretation is just a set of valuations. Intuitively, we can think of this set as representing some constraints on valuations. A valuation v meets these constraints—is *admissible* according to the interpretation—just in case

a) the extension of a expression on v includes all the things that *definitely* fall into the extension of that expression (on the interpretation), and
b) the extensions of expressions on v respect *penumbral connections*.

Penumbral connections are, essentially, relations between the meanings of different vague terms. For example, it might be required that nothing be in the extension of both 'orange' and 'red', and that certain connections hold between the extensions of 'tall' and 'taller than'. We can represent these constraints using *meaning postulates*:

(10) $\forall x \neg(\text{Red}(x) \wedge \text{Orange}(x))$

(11) $\forall x \forall y((\text{Tall}(y) \wedge \text{TallerThan}(x, y)) \supset \text{Tall}(x))$

These are just sentences that we require to be true on all valuations in an interpretation.

Supervaluation is sometimes thought of as a way of cashing out the idea that vagueness arises from "semantic indecision." The classical statement of the idea comes from David Lewis:

> The reason it's vague where the outback begins is not that there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word "outback." Vagueness is semantic indecision. (Lewis 1986, p. 213)

On this way of thinking of things, the admissible valuations represent the different candidate extensions we are "undecided" about.

We can now define truth on an interpretation and validity:

- A sentence is *true on a valuation v* iff it is true in the classical model *v*.
- A sentence is *true on an interpretation I* iff it is *supertrue*—true on every valuation in *I*.
- A sentence is *false on an interpretation I* iff it is *superfalse*—false on every valuation in *I*. (Equivalently, a sentence is false on an interpretation *I* if its negation is true in *I*.)
- An argument is *supervaluationally valid* if every supervaluational interpretation that makes all the premises true makes the conclusion true as well.

Given these definitions, some sentences will be neither true nor false on an interpretation. This means that we give up *bivalence*, the assumption that every sentence is either true or false (on a given interpretation). Consider again our Blob. Our intended interpretation of 'red' and 'pink' doesn't single out classical extensions: there are many admissible valuations. On some of them, Blob will fall into the extension of 'red'; on others, Blob will fall into the extension of 'pink'. So

(12) Blob is red

and

(13) Blob is pink

will be neither true nor false. However, on every admissible valuation, Blob is in either the extension of 'red' or the extension of 'pink', so

(14) Blob is red or pink

comes out true on all admissible valuations, and hence true. We have, then, an example of a true disjunction with neither disjunct true.

Although we reject bivalence on this approach, we get to keep the *Law of Excluded Middle*: each instance of

$$\phi \vee \neg\phi$$

Exercise 8.2: Supervaluationism

Prove that

1. A sentence in the language of first-order logic is true on all classical models if and only if it is true on all supervaluational interpretations.
2. An argument in the language of first-order logic is classically valid if and only if it is supervaluationally valid.

is true on every supervaluational interpretation. (Contrast intuitionism, which gives up both bivalence and the Law of Excluded Middle.)

Indeed, this is just a special case of a more general result. A sentence in the language of first-order logic will be true on all classical models if and only if it is true on all supervaluational interpretations. Similarly, an argument in the language of first-order logic is classically valid if and only if it is supervaluationally valid. (Can you see why?) So we have *classical logic* with a *nonclassical semantics*. Our worries about applying classical semantics to vague language need not call into question the applicability of classical logic.

Still, the supervaluationist's nonclassical semantics is decidedly odd. You might wonder how a sign '$\vee$' can mean *or* if $\ulcorner p \vee q \urcorner$ can be true when neither p nor q is.

8.4.1 Application to sorites

What does supervaluationism say about the sorites paradox?

Recall the basic problem: if we block the paradox by rejecting the universal premise

(15) $\forall n(R(n) \rightarrow R(n-1))$

we seem forced to accept its negation:

(16) $\exists n(R(n) \wedge \neg R(n-1))$.

But this looks like an implausible commitment to a sharp boundary between the rich and the non-rich.

As we have seen, multivalued approaches solve the puzzle by allowing you to reject (15) (or the conditionals that are its instances) as less than fully true, without accepting its negation (16). What does supervaluationism say about the universal premise (15)? Like multivalued approaches, it says you should reject it.

(It's superfalse.) But, unlike multivalued approaches, it says you should accept its negation (16), which is supertrue.

Isn't this just what we found problematic with classical semantics—that rejecting (15) would require accepting (16), a sharp-boundaries claim? So how does supervaluationism improve over classical semantics?

The key difference is that, for the supervaluationist, accepting (16) does not commit one to accepting that there is a true *witness* to the existentially quantified claim: that is, a true instance. In classical semantics, (16) can only be true if there is a true instance of the form

(17) $\exists n(R(n) \wedge \neg R(n-1))$.

But the supervaluationist can deny that any sentence of the form (17) is true, while still accepting (16). (This is just the extension to quantification of the point we already saw: that there can be a true disjunction with neither disjunct true.) This takes the sting out of accepting (16). We can accept that *there is* a number of cents such that having that many cents makes one rich, and having one cent fewer does not make one rich, while rejecting, for every particular number n, the claim that n cents is enough to make you rich while $n-1$ cents is not.

8.4.2 Higher-order vagueness

An admissible valuation, we said, must include all the *definite F*s in the extension of '*F*'. By positing a set of admissible valuations, then, supervaluationism presupposes a sharp boundary between the people who are definitely rich and those who are not. And this, you might think, is just as objectionable as having a sharp boundary between the people who are rich and those who aren't. If the worry was about having sharp, unknowable boundaries, we still have them.

We can sharpen this worry by adding an operator D (for 'definitely'):

$\ulcorner D\phi \urcorner$ is true on a valuation in a supervaluational interpretation I just in case ϕ is true on all valuations in I.

Informally, $\ulcorner D\phi \urcorner$ is true if ϕ is true on every admissible valuation.

With this piece of vocabulary in hand, we can run our sorites paradox with 'definitely enough to make you rich' instead of 'enough to make you rich'. The sorites premise seems almost as compelling:

(18) $\forall n(DR(n) \rightarrow DR(n-1))$
For any n, if n cents is definitely enough to make you rich, then $n-1$ cents is too.

The usual supervaluationist response (e.g., in Keefe 2000, ch. 7) is to say that the notion of admissibility is vague. But if that is right, then we are doing our semantics for vague terms in a vague metalanguage. If we're going to do that, one might wonder, why not just do it at the beginning, and say

(19) The extension of 'rich' is the set of rich things?

Does supervaluational semantics, conducted in a vague metalanguage, have any advantage over classical semantics, conducted in a vague metalanguage?

8.4.3 The logic of definiteness

Williamson (1994, §5.3) calls our attention to some curious facts about the logic of the D operator. The inference schema

(20) $$\frac{\phi}{D\phi}$$

is valid. (Confirm this for yourself.) But

(21) $\phi \supset D\phi$

isn't a logical truth. When ϕ is borderline, (21) is false on some valuations, so it is not supertrue.

So, Williamson observes, once we introduce D into our language, we're going to lose some properties of classical validity—for example,

(22) If ϕ logically implies ψ, then $\phi \supset \psi$ is a logical truth.

(23) If ϕ logically implies ψ, $\neg\psi$ logically implies $\neg\phi$.

(24) If ϕ logically implies ψ and ξ logically implies ψ, then $\ulcorner\phi \vee \xi\urcorner$ logically implies ψ.

How bad are these results? Williamson (1994, p. 152) says that "supervaluations invalidate our natural mode of deductive thinking," such as Conditional Proof (conditional introduction), Dilemma (disjunction elimination), and Reductio Ad Absurdum (negation introduction). But this isn't so clear. In natural deductions for modal logic (§3.1.4), we have already seen that it is possible to restrict what can be done inside a subproof. Could we not impose similar restrictions here? Instead of allowing all valid forms of argument inside subproofs, we could allow only locally valid forms of argument—where an argument is *locally valid* just in case it preserves truth on every valuation on every interpretation (for the distinction between global and local validity, see Williamson 1994, p. 148). This would allow

Exercise 8.3: The logic of D

Provide counterexamples for the inference schemes (23) and (24).

us to use Conditional Proof, Dilemma, and Reductio Ad Absurdum without getting in trouble with the D operator.[6]

8.5 Vagueness in the world?

Recommended reading

Gareth Evans, "Can There Be Vague Objects?" (Evans 1978).

Is vagueness a feature of language? Or is the world itself vague?

Most discussions have assumed the former. Suppose you observe a tower, and, when asked how tall it is, you say "It's pretty tall." You've described its height in a vague way. But of course the tower has a perfectly precise height. It's natural to think, then, that the vagueness here is all in your way of talking, not in the world.

Supervaluationism is explicitly motivated by this thought. The central thought is that the meanings of vague terms are "undecided" between a big cloud of precise denotations. The things that can be denoted are all precise. We get vagueness because it is indeterminate which of these denotations are the denotations of our vague words.

On this way of thinking of things, there are vague *predicates*, but there are no vague *properties or relations*. We have vague *names*, but there are no vague *objects*. There are, to be sure, vague identity statements involving proper names. David Lewis gives the example

(25) Princeton = Princeton Borough,

which is indeterminate because "it is unsettled whether the name 'Princeton' denotes just the Borough, the Borough plus the surrounding Township, or one of countless somewhat larger regions" (Lewis 1988, p. 128). On Lewis's way of thinking, the indeterminacy of (25) just reflects the vagueness of the names flanking the identity.

[6] See McGee and McLaughlin 1998, pp. 224–5, McGee and McLaughlin 2004, and MacFarlane 2008, §4.

But we could try thinking about it another way. We could suppose that, among the properties that things can have, some are vague. On this view, the word 'bald', for example, determinately expresses a single property, the property of being bald. But this is a *vague property*: it is indeterminate exactly which objects have it.

Similarly, among the objects that exist in the world, some are vague. There is no indeterminacy about which object 'Princeton' refers to—it refers to Princeton—but this is a vague object. It is indeterminate whether this object is identical to Princeton Borough or to various other regions that might be picked out in other ways.

A nice motivating example is the ship of Theseus. Recall that the planks in the original ship were replaced one by one over time and the old planks stored in a warehouse. Eventually the ship of Theseus contains none of its old wood, but this does not prevent it from being the same ship as the original. At this point, however, the old planks are taken from the warehouse and used to build a ship using the original plan. The new ship has both the form and the material of the original ship, so it, too, has a claim to be the original ship of Theseus. It is tempting to say that it is indeterminate which of these ships is identical with the original ship of Theseus. The indeterminacy does not seem to arise from language: we can raise the question no matter what name or description we use to pick out the original ship and the new one.

8.5.1 Evans on vague identity

Can we really understand what it would be for it to be indeterminate whether two objects are the same? Evans (1978) thinks that the idea is logically incoherent.

In his argument, Evans uses

'∇' to mean *it is indefinite whether*, and

'Δ' to mean *it is definite whether*.

We will instead use

'Δ' to mean *it is definite that*, and

'∇' to mean *it is not definite that not*.

This makes '∇' and 'Δ' behave like '$\Box$' and '$\Diamond$' in modal logic. (On this understanding, unlike Evans', $\ulcorner \Delta\, \phi \urcorner$ entails ϕ and $\ulcorner \nabla\, \phi \urcorner$.)

Evans also uses the notation '$\hat{x}\phi x$' to mean *the property of being* ϕ. (You may have seen a different notation for this: '$\lambda x(\phi x)$'.) But we can present the argument without this notation.

***Exercise 8.4: The logic of* Δ *and* ∇**

Evans' proof depends on the logic of 'Δ' and '∇' being S5. Is this a plausible assumption? Can you think of reasons for or against it? Does another modal logic seem more appropriate for the idea of "definiteness"?

The core argument is as follows:

1	$\neg \Delta a{=}b \wedge \neg \Delta \neg a{=}b$	Hyp	
2	$\Delta a{=}a$	assumption	
3	$\neg \Delta a{=}b$	$\wedge$ Elim 1	
4	$a{=}b$	Hyp	(8.1)
5	$\Delta a{=}b$	= Elim + Reit, 2, 4	
6	$\bot$	$\neg$ Elim + Reit 3, 5	
7	$\neg a{=}b$	$\neg$ Intro 4–6	

In line 1 we assume that it's indefinite *whether* $a{=}b$: that is, it is not definite that $a{=}b$, and not definite that $\neg a{=}b$. What Evans wants to show is that this assumption, which should be coherent if there are vague objects, is incoherent. To do that, he draws on the assumption that it is definite that $a{=}a$ (line 2). He then argues that $\neg a{=}b$. The idea is simple: if we had $a{=}b$, then we would be able to substitute 'b' for 'a' in '$\Delta\, a{=}a$' to get '$\Delta\, a{=}b$', and this would contradict our hypothesis (1). (Note that our = Elim rule plays the role played in Evans' argument by "Leibniz's Law.") So, $\neg a{=}b$.

Having given this argument, Evans cheats a bit. He says that this conclusion "contradict[s] the assumption, with which we began, that the identity statement '$a{=}b$' is of indeterminate truth value." But it doesn't, exactly. What would contradict hypothesis 1 is '$\Delta \neg a{=}b$', but we only have '$\neg a{=}b$'. Evans is aware of this. He says that we can get a real contradiction with the assumption that his operators "generate a modal logic as strong as S5." Though he does not elaborate, we can see from the following Fitch proof that what he says is essentially correct. Here we assume that our 'Δ' and '∇' operators work like S5 '$\Box$' and '$\Diamond$'.

1	$\neg \Delta\, a{=}b \wedge \neg \Delta \neg a{=}b$	Hyp	
2	$\neg \Delta\, a{=}b$	$\wedge$ Elim 1	
3	$\nabla \neg a{=}b$	MNE 2	
4	Δ | $a{=}a$	= Intro	
5	$\Delta\, a{=}a$	Δ Intro 4	
6	Δ | $a{=}b$	Hyp	
7	$\nabla \neg a{=}b$	Modal Reit S5, 3	
8	$\neg \Delta\, a{=}b$	MNE 7	(8.2)
9	$\Delta\, a{=}a$	Modal Reit S4, 5	
10	$\Delta\, a{=}b$	= Elim 6, 9	
11	$\bot$	$\neg$ Elim 8, 10	
12	$\neg a{=}b$	$\neg$ Intro 6–11	
13	$\Delta \neg a{=}b$	Δ Intro 6–12	
14	$\neg \Delta \neg a{=}b$	$\wedge$ Elim 1	
15	$\bot$	$\neg$ Elim 13, 14	

We have reduced to absurdity the assumption that it is indefinite whether $a{=}b$.

8.5.2 Evans and Quine

If the form of Evans' argument looks familiar, it's for good reason. Recall Quine's argument against quantified modal logic (§3.2). One version goes like this:

1	$\neg\Box a{=}b$	Hyp	
2	$\Box a{=}a$	assumption	
3	| $a{=}b$		(8.3)
4	| $\Box a{=}b$	= Elim 2, 3	
5	| $\bot$	$\neg$ Elim + Reit 1, 4	
6	$\neg a{=}b$	$\neg$ Intro 2–5	

This is essentially the same as Evans' core argument (8.1), with '□' in place of 'Δ'.[7]

We looked at two different possible responses to Quine's argument. One response—Quine's—is to conclude that *de re* necessity and possibility do not make sense. Being necessarily identical to *a* isn't a property *a* has, so we cannot conclude that if *a*=*b*, *b* has the same property. Talk of necessity and possibility has to do not with how things are in the world, but with how we describe them.

The other response—Kripke's—is to take the conclusion of the argument at face value: if *a*=*b*, then □*a*=*b*. That is, there are no contingent relations of identity. An object can have many properties contingently—it can be blue contingently, for example—but it cannot be contingently identical to another object. Relations of identity are necessary.

What would parallel responses to Evans' argument look like?

The parallel to Quine's response would be to claim that *de re* definiteness does not make sense. Being definitely identical to *a* isn't a property *a* has, so we cannot conclude that if *a*=*b*, *b* has the same property. Talk of definiteness has to do not with how things are in the world, but with how we describe them. This response is a decisive rejection of the idea that there can be "vagueness in the world."

The parallel to Kripke's response is to accept the conclusion of the argument and hold that, if *a*=*b*, then Δ *a*=*b*. That is, there are no indefinite relations of identity. This response does not amount to a rejection of the idea that there can be vagueness in the world. After all, although Kripke didn't think *being identical with Nixon* was a property Nixon had contingently, he did think that *winning the 1972 US Presidential election* was a property Nixon had contingently. Similarly, a friend of worldly vagueness could still hold that there are many properties and relations, other than relations of identity, that have indefinite extensions. These could include properties like *being bald* and relations like *is a material part of*.

Still, one might object: isn't it obvious that there are indefinite statements of identity? For example, plausibly,

(26) ∇ (Theseus's original ship = the ship just built from the same wood to the same plan).

But the truth of (26) is no more a threat to the view that relations of identity hold definitely than the truth of

(27) ◊ (the number of planets = 4)

is to the view that relations of identity hold necessarily. For, as we saw in Chapter 3, 'the number of planets' is not a rigid designator: it denotes different objects in different possible worlds. Similarly, as long as the descriptions in (26) are not what

[7] The analogy to the modal case is noted by Thomason (1982).

Thomason 1982 calls *precise designators*—terms that denote the same object no matter how vagueness is resolved—(26) can be true even if identity always holds definitely between objects.

Are there any indefinite identity statements involving only precise designators? Lewis's example might seem a good candidate:

(28) ∇ Princeton = Princeton Borough.

If all proper names are rigid designators, then Evans' argument gives us a reason to reject (28). But are all proper names precise designators? We can accept (28) if we hold that 'Princeton' is not a precise designator (and should not be represented by an individual constant in a logical system containing = Elim).

What Evans' argument does, plausibly, show (as Lewis observes) is that even if there is vagueness in the world, it cannot explain the truth of sentences like (28). For that, we need at least some vagueness to depend on our mode of describing the world.

Further readings

- Sainsbury 1995, ch. 2 is a nice informal introduction to vagueness and the sorites paradox.
- The anthology Keefe and Smith 1997 contains an excellent selection of classic papers on the logic and semantics of vagueness.
- For comprehensive surveys of the main theories, see Williamson 1994 and Keefe 2000: Williamson favors epistemicism, while Keefe argues for super-valuationism.
- Since these books appeared, there has been more attention to the ways in which vague terms are context-sensitive (a topic not covered here). See Soames 1999, ch. 7, Fara 2000, Shapiro 2006, Kennedy 2007, and MacFarlane 2016.

Appendix A Greek Letters

Here is a pronunciation guide for the Greek letters used in this book (so you don't have to say 'squiggle').

Lowercase	Uppercase	Name	Pronunciation
α	A	alpha	alfa
β	B	beta	bayta
ϕ	Φ	phi	fee or fie
ψ	Ψ	psi	psee or psigh or see or sigh
ξ	Ξ	xi	ksee or ksigh
ρ	P	rho	row
π	Π	pi	pee or pie
σ	Σ	sigma	sigma

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Appendix B Set-Theoretic Notation

Table B.1 is a guide to the set-theoretic notation we use in this book. You only need to know the very basics of set theory in order to understand this notation.

Sets are collections of things. The things that belong to a set are its members. To indicate that something is a member of a set, we use the symbol '$\in$'. For example, '$a \in S$' means 'a is a member of the set S'.

Sets are individuated by their members, so if S and T have the same members, they are the same set. There is a set with no members, called the *empty set* or '$\emptyset$'.

The members of a set may be anything: ordinary objects like books and tables, abstract objects like numbers, even other sets. The set $\{\emptyset\}$, for example, is the set whose sole member is the empty set. Note that 1 is a member of $\{1\}$ but not of $\{\{1\}\}$.

A set's *cardinality* is the number of members it has. Two sets have the same cardinality if their members can be matched up one-to-one. Some sets (for example, the set of natural numbers) have infinitely many members. Most of the interesting results of set theory have to do with sets of infinite cardinality. For example, it can be proven that the cardinality of the set of real numbers is greater than the cardinality of the set of natural numbers, even though they are both infinite, and that the cardinalities of the set of even natural numbers and the set of natural numbers are the same, even though the former set is a proper subset of the latter. But these results are beyond the scope of this primer, and aren't necessary for this book.

Ordered pairs (and ordered n-tuples generally) are written using angle brackets: for example, $\langle 1, 2\rangle$ is the ordered pair consisting of 1 and 2, in that order, and $\langle 1, 2, 3\rangle$ is the ordered triple consisting of 1, 2, and 3, in that order. Ordered pairs can be defined in terms of sets, as follows:

$$\langle x, y\rangle =_{def} \{\{x\}, \{x, y\}\}$$

Note that $\{1, 2\} = \{2, 1\}$, but $\langle 1, 2\rangle \neq \langle 2, 1\rangle$.

Notation	Meaning
$\{1, 2, 3\}$	the set whose members are 1, 2, and 3
$\{x : \phi x\}$	the set of things that satisfy ϕx
	for example, $\{x : x < 3 \wedge x \in \mathbb{N}\} = \{0, 1, 2\}$
$x \in S$	x is a member of S
$x \notin S$	x is not a member of S
$\emptyset$	the empty set
$\lvert S \rvert$	the cardinality of S
$S \cap T$	the *intersection* of S and T
	(the set of things that are members of both S and T)
$S \cup T$	the *union* of S and T
	(the set of things that are members of either S or T)
$S - T$	the *difference* of S and T
	(the set containing all the members of S that are *not* in T)
$S \subseteq T$	S is a *subset* of T
	(every member of S is a member of T)
$S \subset T$	S is a *proper subset* of T
	(S is a subset of T but is not identical to T))
$\langle 1, 2 \rangle$	the ordered pair whose first member is 1 and whose second member is 2

Table B.1: Set-theoretic notation.

Appendix C Proving Unrepresentability

Boolos (1984, p. 57) offers a proof (due to David Kaplan) that

(B) $\exists X(\exists x Xx \wedge \forall x \forall y[(Xx \wedge Axy) \supset (x \neq y \wedge Xy)])$

cannot be given a first-order formulation. For those who have studied some metalogic, we walk through the proof here. In broad outline, it runs as follows:

1. If there were a first-order formula that captured the meaning of (B), it would be possible to give first-order axioms for arithmetic that rule out nonstandard models.

2. But it can be proven that no first-order axioms for arithmetic can rule out nonstandard models.

3. Hence (by reductio) there is no first-order formula that captures the meaning of (B).

To understand this, we need to know what a *nonstandard model of arithmetic* is. You already know what a *model* is: a domain and an interpretation of the language's predicates and individual constants on that domain. A model of a set of axioms is a model that makes these axioms true. Now consider a set of first-order axioms for arithmetic (such as the standard Peano axioms). These axioms will contain some arithmetical expressions, like '0', '*S*', '+', and '<'. The *standard model* of arithmetic interprets these in the normal way: the domain is the set of natural numbers, the extension of '*S*' is the set of pairs consisting of a natural number and its successor, the extension of '+' is the set of triples consisting of two natural numbers and their sum, and the extension of '<' is the set of pairs consisting of two natural numbers where the first is less than the second.

Surprisingly, though, the standard model is not the only model of the axioms. There are also *nonstandard models* whose domains contain lots of "extra numbers" that are greater than all the standard natural numbers. In a nonstandard model, the number series looks like this:

$$\underbrace{0, 1, 2, 3, 4, 5, 6, 7, 8, \ldots}_{\text{all the standard natural numbers}} \qquad a, a + 1, a + 2, \ldots$$

These nonstandard numbers a, $a + 1$, ... are numbers that you could never get to by starting with 0 and moving in a finite number of steps to the next number.

Here is a proof of the existence of nonstandard models. The *compactness theorem* for first-order logic says that a set of sentences has a model iff every finite subset of the set has a model. Let A be your first-order axioms for arithmetic, and consider the set $S{=}A \cup \{Na, a{\neq}0, a{\neq}1, a{\neq}2, \cdots\}$. Clearly, any finite subset T of this set has a model—just interpret a as the smallest natural number not mentioned in T. So, by compactness, the whole set S has a model. In this model, Na is true but a cannot denote any of the standard numbers.

You might think that the principle of mathematical induction rules out nonstandard numbers. In its natural second-order formulation

$$\forall X([X0 \wedge \forall x \forall y((Xx \wedge Syx) \supset Xy)] \supset \forall x(Nx \supset Xx))$$

the principle says that any property that belongs to 0 and belongs to the successor of a number if it belongs to that number, belongs to all natural numbers. How could that be true if there are nonstandard numbers that can never be reached by starting with 0 and moving to the successor? But remember, in first order logic we just have an induction *schema*,

$$[\phi 0 \wedge \forall x \forall y((\phi x \wedge Syx) \supset \phi y)] \supset \forall x(Nx \supset \phi x).$$

This ensures that any property that is *expressible in the language*, belongs to 0, and belongs to the successor of a number if it belongs to that number, belongs to all natural numbers. This might be true even if there are properties, inexpressible in the language, that belong to the standard natural numbers but not the nonstandard ones.

Kaplan establishes premise 1 of his argument by giving a sentence (C) that is a substitution instance of (B), with '$(x{=}0 \vee x{=}y + 1)$' put in for 'Axy'. Clearly, if there is a first-order representation of (B), there will be a first-order representation of (C). He then shows that (C) is true in every nonstandard model of arithmetic, but false in the standard model. To see this, define '*x doodles y*' as '$x{=}0 \vee x{=}y + 1$', so that 0 doodles everything and other numbers doodle their predecessors. Then (C) says:

(1) There are some numbers that only doodle each other.

(1) will be false in standard models of arithmetic. Consider any group of standard numbers. If it contains 0, then it can't be a group of numbers that only doodle each other, since 0 doodles itself. If it doesn't contain 0, then let k be the least number it contains. Since k is the least number in the group, $k - 1$ is not in the group. But k doodles $k - 1$. So again, it can't be a group of numbers that only doodle each other.

But (1) is true in all nonstandard models of arithmetic, since these models contain infinite descending chains of numbers that don't bottom out in 0.

Thus, if there were a first-order formula equivalent to (C), we would have a first-order way to rule out all nonstandard models of arithmetic: just add the negation of (C) to the other axioms. Since it can be proven on general grounds that there is no way to do this, we know there can't be a first-order formula equivalent to (C).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

References

Ackermann, Willhelm (1956). "Begründung einer strengen Implikation." *Journal of Symbolic Logic* 21, pp. 113–128.

Anderson, Alan Ross and Nuel D. Belnap Jr. (1975). *Entailment*. Vol. 1. Princeton: Princeton University Press.

Anderson, Alan Ross, Nuel D. Belnap Jr., and J. Michael Dunn (1992). *Entailment*. Vol. 2. Princeton: Princeton University Press.

Aristotle (1984). *The Complete Works of Aristotle. The Revised Oxford Translation*. Ed. by Jonathan Barnes. Vol. 1. Princeton: Princeton University Press.

Barwise, Jon and Robin Cooper (1981). "Generalized Quantifiers and Natural Language." *Linguistics and Philosophy* 4, pp. 159–219.

Barwise, Jon and John Etchemendy (1999). *Language, Truth and Logic*. Palo Alto, CA: CSLI.

Barwise, Jon and Solomon Feferman (1985). *Model-Theoretic Logics*. Berlin: Springer.

Barwise, Jon and John Perry (1981). "Semantic Innocence and Uncompromising Situations." *Midwest Studies in Philosophy* 6, pp. 387–404.

Beall, Jc and Greg Restall (2006). *Logical Pluralism*. Oxford: Oxford University Press.

Belnap Jr., Nuel D. (1961). "Tonk, Plonk and Plink." *Analysis* 22, pp. 130–134.

Belnap Jr., Nuel D. (2009). "Notes on the Art of Logic." URL: `http://www.pitt.edu/~belnap/nal.pdf`.

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bolzano, Bernard (1929/1931). *Wissenschaftslehre. Versuch einer ausführlichen und grösstentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter*. 2nd ed. Leipzig: Felix Meiner.

Boolos, George (1975). "On Second-Order Logic." *Journal of Philosophy* 72, pp. 509–527.

Boolos, George (1984). "To Be is to Be a Value of a Variable (or to Be Some Values of Some Variables)." *Journal of Philosophy* 81, pp. 430–449.

Boolos, George (1985). "Nominalist Platonism." *Philosophical Review* 94, pp. 327–344.

Burgess, John (2005). "No Requirement of Relevance." In: *The Oxford Handbook of Philosophy of Mathematics and Logic*. Ed. by Stewart Shapiro. Oxford: Oxford University Press, pp. 727–750.

Burgess, John (2009). *Philosophical Logic*. Princeton: Princeton University Press.

Carnap, Rudolf (1956). *Meaning and Necessity*. 2nd ed. Chicago: University of Chicago Press.

Carnap, Rudolf (2002). *The Logical Syntax of Language*. Trans. by A. Smeaton. Open Court Classics. Peru, IL: Open Court.
Christensen, David (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford: Oxford University Press.
Church, Alonzo (1943). "Review of R. Carnap, *Introduction to Semantics*." *Philosophical Review* 52, pp. 298–304.
Church, Alonzo (1956). *Introduction to Mathematical Logic*. 2nd ed. Princeton: Princeton University Press.
Coffa, J. Alberto (1975). "Machian Logic." *Communication and Cognition* 8, pp. 103–129.
Coffa, J. Alberto (1991). *The Semantic Tradition from Kant to Carnap*. Cambridge: Cambridge University Press.
Cook, Roy T. (2010). "Let a Thousand Flowers Bloom: A Tour of Logical Pluralism." *Philosophy Compass* 5, pp. 492–504.
Correia, Fabrice (2003). "Review of Stephen Neale, *Facing Facts*." *Dialectica* 57, pp. 439–444.
Davidson, Donald (1984). "True to the Facts." In: *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.
Dummett, Michael (1964). "Bringing About the Past." *Philosophical Review* 73, pp. 338–359.
Dummett, Michael (1978). "The Philosophical Basis of Intuitionistic Logic." In: *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press, pp. 215–247.
Dummett, Michael (1979). *Elements of Intuitionism*. Oxford: Oxford University Press.
Dummett, Michael (1991). *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.
Dunn, J. Michael and Nuel D. Belnap Jr. (1968). "The Substitution Interpretation of the Quantifiers." *Nous* 2, pp. 177–185.
Dunn, J. Michael and Greg Restall (2002). "Relevance Logic." In: *Handbook of Philosophical Logic*. Ed. by D. Gabbay and F. Guenthner. 2nd ed. Vol. 6. Dordrecht: Springer, pp. 1–128.
Edgington, Dorothy (1993). "Do Conditionals Have Truth-Conditions?" In: *A Philosophical Companion to First-Order Logic*. Ed. by R. I. G. Hughes. Indianapolis: Hackett, pp. 28–49.
Edgington, Dorothy (1997). "Vagueness by Degrees." In: *Vagueness: A Reader*. Ed. by Rosanna Keefe and Peter Smith. Cambridge, MA: MIT, pp. 294–316.
Edgington, Dorothy (2014). "Indicative Conditionals." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2014. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/win2014/entries/conditionals/`.
Etchemendy, John (1990). *The Concept of Logical Consequence*. Cambridge, MA: Harvard University Press.
Etchemendy, John (2008). "Reflections on Consequence." In: *New Essays on Tarski and Philosophy*. Ed. by Douglas Patterson. Oxford: Oxford University Press.
Evans, Gareth (1978). "Can There Be Vague Objects?" *Analysis* 38, p. 208.

Fara, Delia Graff (2000). "Shifting Sands: An Interest-Relative Theory of Vagueness." *Philosophical Topics* 28, pp. 45–81.
Field, Hartry (2000). "Indeterminacy, Degree of Belief, and Excluded Middle." *Nous* 34, pp. 1–30.
Field, Hartry (2009a). "Pluralism in Logic." *Review of Symbolic Logic* 2, pp. 342–359.
Field, Hartry (2009b). "The Normative Role of Logic." *Proceedings of the Aristotelian Society* s.v. 83, pp. 251–268.
Fine, Kit (1997). "Vagueness, Truth and Logic." In: *Vagueness: A Reader*. Ed. by Rosanna Keefe and Peter Smith. Cambridge, MA: MIT, pp. 119–150.
Fitch, Frederic Brenton (1952). *Symbolic Logic: An Introduction*. New York: Ronald Press Co.
Foley, Richard (1992). "The Epistemology of Belief and the Epistemology of Degrees of Belief." *American Philosophical Quarterly* 29, pp. 111–124.
Forbes, Graham (1985). *The Metaphysics of Modality*. Oxford: Oxford University Press.
Frege, Gottlob (1892). "Über Sinn und Bedeutung." *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50.
Frege, Gottlob (1893). *Grundgesetze der Arithmetik*. Vol. 1. Jena: H. Pohle.
Frege, Gottlob (1980). "On Sense and Meaning." In: *Translations from the Philosophical Writings of Gottlob Frege*. Ed. and trans. by Peter Geach and Max Black. 3rd ed. Oxford: Blackwell, pp. 56–78.
Geach, P. T. (1958). "Entailment." *Proceedings of the Aristotelian Society* s.v. 32, pp. 157–172.
Geach, P. T. (1970). "Entailment." *Philosophical Review* 79, pp. 237–239.
Gibbard, Allan (1981). "Two Recent Theories of Conditionals." In: *IFS: Conditionals, Belief, Decision, Chance, and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce, pp. 211–247.
Gibbard, Allan (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.
Girle, Rod (2000). *Modal Logics and Philosophy*. Montreal: McGill-Queen's.
Gödel, Kurt (1944). "Russell's Mathematical Logic." In: *The Philosophy of Bertrand Russell*. Ed. by P. A. Schilpp. Evanston: Northwestern University Press, pp. 125–153.
Goguen, J. A. (1969). "The Logic of Inexact Concepts." *Synthese* 19, pp. 325–73.
Gómez-Torrente, Mario (1996). "Tarski on Logical Consequence." *Notre Dame Journal of Formal Logic* 37, pp. 125–151.
Gómez-Torrente, Mario (2002). "The Problem of Logical Constants." *Bulletin of Symbolic Logic* 8, pp. 1–37.
Goodman, Nelson (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.
Grice, H. P. (1989). "Logic and Conversation." In: *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Harman, Gilbert (1984). "Logic and Reasoning." *Synthese* 60, pp. 107–127.
Heyting, Arend (1956). *Intuitionism: An Introduction*. Amsterdam: North Holland Publishing.

Hughes, G. E. and M. J. Cresswell (1996). *A New Introduction to Modal Logic*. London: Routledge.

Kant, Immanuel (1965). *Critique of Pure Reason*. Trans. by Norman Kemp Smith. New York: St. Martin's Press.

Katz, Bernard D. (1999). "On a Supposed Counterexample to Modus Ponens." *Journal of Philosophy* 96, pp. 404–415.

Keefe, Rosanna (2000). *Theories of Vagueness*. Cambridge: Cambridge University Press.

Keefe, Rosanna and Peter Smith, eds. (1997). *Vagueness: A Reader*. Cambridge, MA: MIT.

Kennedy, Christopher (2007). "Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives." *Linguistics and Philosophy* 30, pp. 1–45.

Kleene, Stephen Cole (1952). *Introduction to Metamathematics*. New York: Van Nostrand.

Kolodny, Niko and John MacFarlane (2010). "Ifs and Oughts." *Journal of Philosophy* 107, pp. 115–143.

Kripke, Saul (1965). "Semantical Analysis of Intuitionistic Logic." In: *Formal Systems and Recursive Functions*. Ed. by J. Crossley and M. A. E. Dummett. Amsterdam: North Holland Publishing, pp. 92–130.

Kripke, Saul (1976). "Is There a Problem about Substitutional Quantification?" In: *Truth and Meaning*. Ed. by Gareth Evans and John McDowell. Oxford: Oxford University Press.

Kripke, Saul (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, C. I. and C. H. Langford (1959). *Symbolic Logic*. 2nd ed. New York: Dover.

Lewis, David (1968). "Counterpart Theory and Quantified Modal Logic." *Journal of Philosophy* 65, pp. 113–126.

Lewis, David (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, David (1976). "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85, pp. 297–315.

Lewis, David (1986). *On the Plurality of Worlds*. Oxford: Basil Blackwell.

Lewis, David (1988). "Vague Identity: Evans Misunderstood." *Analysis* 48, pp. 128–130.

Lewis, David (1998). "Logic for Equivocators." In: *Papers in Philosophical Logic*. Cambridge: Cambridge University Press, pp. 97–110.

Lindström, Per (1966). "First Order Predicate Logic with Generalized Quantifiers." *Theoria* 32, pp. 186–195.

Linsky, Bernard (2016). "The Notation in Principia Mathematica." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2016. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/fall2016/entries/pm-notation/`.

Linsky, Leonard (1972). "Two Concepts of Quantification." *Nous* 6, pp. 224–239.

Loux, Michael J., ed. (1979). *The Possible and the Actual*. Ithica, NY: Cornell University Press.

MacFarlane, John (2006). "The Things We (Sorta Kinda) Believe." *Philosophy and Phenomenological Research* 73, pp. 218–224.

MacFarlane, John (2008). "Truth in the Garden of Forking Paths." In: *Relative Truth*. Ed. by Max Kölbel and Manuel Garcia-Carpintero. Oxford: Oxford University Press, pp. 81–102.
MacFarlane, John (2010). "Fuzzy Epistemicism." In: *Cuts and Clouds*. Ed. by Richard Dietz and Sebastiano Moruzzi. Oxford: Oxford University Press, pp. 438–463.
MacFarlane, John (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.
MacFarlane, John (2016). "Vagueness as Indecision." *Proceedings of the Aristotelin Society* s.v. 90, pp. 255–283.
MacFarlane, John (2017). "Logical Constants." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/win2017/entries/logical-constants/`.
Marcus, Ruth Barcan (1962). "Interpreting Quantification." *Inquiry* 5, pp. 252–259.
Marcus, Ruth Barcan (1972). "Quantification and Ontology." *Nous* 6, pp. 240–250.
McGee, Vann (1985). "A Counterexample to Modus Ponens." *Journal of Philosophy* 82, pp. 462–471.
McGee, Vann and Brian McLaughlin (1998). "Review of Timothy Williamson, *Vagueness*." *Linguistics and Philosophy* 21, pp. 221–235.
McGee, Vann and Brian McLaughlin (2004). "Logical Commitment and Semantic Indeterminacy: A Reply to Williamson." *Linguistics and Philosophy* 27, pp. 221–235.
Meyer, Robert K. (1971). "Entailment." *Journal of Philosophy* 68, pp. 808–818.
Mostowski, Andrzej (1957). "On a Generalization of Quantifiers." *Fundamenta Mathematicae* 44, pp. 12–36.
Neale, Stephen (1990). *Descriptions*. Cambridge, MA: MIT Press.
Neale, Stephen (1995). "The Philosophical Significance of Gödel's Slingshot." *Mind* 104, pp. 761–825.
Neale, Stephen (2001). *Facing Facts*. Oxford: Oxford University Press.
Normore, Calvin (1993). "The Necessity in Deduction: Cartesian Inference and Its Medieval Background." *Synthese* 96, pp. 437–454.
Ostertag, Gary, ed. (1998). *Definite Descriptions: A Reader*. Cambridge, MA: MIT Press.
Parry, W. T. (1932). "Implication." PhD thesis. Harvard University.
Parry, W. T. (1933). "Ein Axiomensystem für eine neue Art von Implikation (analytische Implikation)." *Ergebnisse eines mathematischen Kolloquiums* 4, pp. 4–6.
Platts, Mark (1979). *Ways of Meaning*. London: Routledge and Kegan Paul.
Prawitz, Dag (1985). "Remarks on Some Approaches to the Concept of Logical Consequence." *Synthese* 62, pp. 153–171.
Prawitz, Dag (2005). "Logical Consequence From a Constructivist Point of View." In: *The Oxford Handbook of Philosophy of Mathematics and Logic*. Ed. by Stewart Shapiro. Oxford: Oxford University Press, pp. 671–695.
Prawitz, Dag (2006). "Meaning Approached via Proofs." *Synthese* 148, pp. 507–524.
Price, Huw (1983). "Does 'Probably' Modify Sense?" *Australasian Journal of Philosophy* 61, pp. 396–408.

Priest, Graham (1979). "Two Dogmas of Quineanism." *Philosophical Quarterly* 29, pp. 289–301.
Priest, Graham (1998). "What Is So Bad About Contradictions?" *Journal of Philosophy* 95, pp. 410–426.
Prior, A. N. (1960). "The Runabout Inference-Ticket." *Analysis* 21, pp. 38–39.
Puryear, Stephen (2013). "Frege on Vagueness and Ordinary Language." *Philosophical Quarterly* 250, pp. 120–140.
Putnam, Hilary (1968). "Is Logic Empirical?" *Boston Studies in the Philosophy of Science* 5, pp. 216–241.
Quine, W. V. O. (1940). *Mathematical Logic*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. (1948). "On What There Is." *Review of Metaphysics* 2, pp. 21–38.
Quine, W. V. O. (1951). "Two Dogmas of Empiricism." *Philosophical Review* 60, pp. 20–43.
Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. O. (1961). "Reference and Modality." In: *From a Logical Point of View*. 2nd ed. Cambridge, MA: Harvard University Press, pp. 139–159.
Quine, W. V. O. (1970). *Philosophy of Logic*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. (1976). "Three Grades of Modal Involvement." In: *The Ways of Paradox and Other Essays*. Revised and enlarged. Cambridge, MA: Harvard University Press, pp. 158–176.
Ray, Grey (1996). "Logical Consequence: A Defense of Tarski." *Journal of Philosophical Logic* 25, pp. 617–677.
Rayo, Agustín and Stephen Yablo (2001). "Nominalism through De-Nominalization." *Nous* 35, pp. 74–92.
Read, Stephen (1988). *Relevant Logic: A Philosophical Examination of Inference*. Oxford: Basil Blackwell.
Read, Stephen (1994). "Formal and Material Consequence." *Journal of Philosophical Logic* 23, pp. 247–265.
Rieger, Adam (2013). "Conditionals are Material: The Positive Arguments." *Synthese* 190, pp. 3161–3174.
Russell, Bertrand (1905). "On Denoting." *Mind* 14, pp. 479–493.
Russell, Bertrand (1920). *Introduction to Mathematical Philosophy*. 2nd ed. London: George Allen and Unwin.
Russell, Bertrand and Alfred North Whitehead (1910). *Principia Mathematica*. Vol. 1. Cambridge: Cambridge University Press.
Sagüillo, José M. (1997). "Logical Consequence Revisited." *Bulletin of Symbolic Logic* 3, pp. 216–241.
Sainsbury, R. M. (1995). *Paradoxes*. 2nd ed. Cambridge: Cambridge University Press.
Schiffer, Stephen (2003). *The Things We Mean*. Oxford: Oxford University Press.
Shapiro, Stewart (1991). *Foundations without Foundationalism: A Case for Second-order Logic*. Oxford: Clarendon Press.
Shapiro, Stewart (2006). *Vagueness in Context*. Oxford: Oxford University Press.
Shapiro, Stewart (2014). *Varieties of Logic*. Oxford: Oxford University Press.

Sher, Gila (1991). *The Bounds of Logic: A Generalized Viewpoint*. Cambridge, MA: MIT Press.
Sher, Gila (1996). "Did Tarski Commit 'Tarski's Fallacy'?" *Journal of Symbolic Logic* 61, pp. 653–686.
Smiley, Timothy (1959). "Entailment and Deducibility." *Proceedings of the Aristotelian Society* n.s. 59, pp. 233–254.
Smullyan, Arthur (1948). "Modality and Description." *Journal of Symbolic Logic* 13, pp. 31–37.
Soames, Scott (1999). *Understanding Truth*. Oxford: Oxford University Press.
Stalnaker, Robert (1975). "Indicative Conditionals." *Philosophia* 5, pp. 269–286.
Stalnaker, Robert (1976). "Possible Worlds." *Nous* 10, pp. 65–75.
Steinberger, Florian (2017). "The Normative Status of Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University. URL: `https://plato.stanford.edu/archives/spr2017/entries/logic-normative/`.
Steinberger, Florian (2019). "Logical Pluralism and Logical Normativity." *Philosopher's Imprint* 19, pp. 1–19.
Stevenson, J. T. (1961). "Roundabout the Runabout Inference-ticket." *Analysis* 21, pp. 124–128.
Strawson, P. F. (1958). "Review of G. H. von Wright, *Logical Studies*." *Philosophical Quarterly* 8, pp. 372–376.
Tarski, Alfred (1935). "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica* 1, pp. 261–405.
Tarski, Alfred (1983a). "On the Concept of Logical Consequence." In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. Indianapolis: Hackett, pp. 409–420.
Tarski, Alfred (1983b). "The Concept of Truth in Formalized Languages." In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. Indianapolis: Hackett, pp. 152–278.
Tarski, Alfred (1986). "What are Logical Notions?" *History and Philosophy of Logic* 7. Ed. by John Corcoran, pp. 143–154.
Tennant, Neil (1994). "The Transmission of Truth and the Transitivity of Deduction." In: *What Is a Logical System?* Ed. by D. M. Gabbay. Oxford: Oxford University Press, pp. 161–177.
Thomason, Richmond H. (1982). "Identity and Vagueness." *Philosophical Studies* 42, pp. 329–332.
Thomson, James F. (1990). "In Defense of '⊃'." *Journal of Philosophy* 87, pp. 57–70.
Van Inwagen, Peter (1981). "Why I Don't Understand Substitutional Quantification." *Philosophical Studies* 39, pp. 281–285.
Weatherson, Brian (2005). "True, Truer, Truest." *Philosophical Studies* 123, pp. 47–70.
White, Morton (1987). "A Philosophical Letter of Alfred Tarski." *Journal of Philosophy* 84, pp. 28–32.
Williamson, Timothy (1987). "Equivocation and Existence." *Proceedings of the Aristotelian Society* n.s. 88, pp. 109–127.

Williamson, Timothy (1994). *Vagueness*. London: Routledge.
Wittgenstein, Ludwig (1958). *Philosophical Investigations*. 2nd ed. Oxford: Blackwell.
Wolfram, Sybil (1989). *Philosophical Logic*. London: Routledge.
Wright, Crispin (2001). "On Being in a Quandary: Relativism, Vagueness, Logical Revisionism." *Mind* 110, pp. 45–98.
Wright, G. H. von (1957). *Logical Studies*. London: Routledge.

Index