Philosophical Logic

Lecture Notes

Draft from April 12, 2024.
Please don't cite or distribute without permission.
Comments welcome!

Levin Hornischer
Levin.Hornischer@lmu.de

Contents

Preface

These are the lecture notes for the course "Philosophical Logic" given during the summer semester 2024 at *LMU Munich* as part of the *Master in Logic and Philosophy of Science*. Previous editions of the course were given at *LMU Munich* and the *University of Amsterdam*. A website for the course is found here:

https://levinhornischer.github.io/PhilLogic/.

Disclaimer These notes are not a polished and thoroughly checked textbook, but they are intended as hopefully helpful material to complement the lectures.

Comments I'm happy about any comments: spotting typos, finding mistakes, pointing out parts that were confusing on a first read or need better explanation, or simply questions triggered by the material. Just send an informal email to Levin.Hornischer@lmu.de.

Motivation Often, what is understood by 'logic' is classical logic (either classical propositional logic or classical first-order logic). Its typical features are, for example, that a sentence p is either true or false and that 'not not p' is the same as 'p'. In many situations, though, simply assuming classical logic is not appropriate. Such situations arise, as we'll see, in philosophy and neighboring disciplines (like mathematics, computer science, and linguistics). For example, when reasoning with

1. vague concepts
2. incomplete or inconsistent data
3. conditionals (i.e., various 'if, then' sentences)
4. what it means to be true.

The course will be about exploring different logics to model these (and other) phenomena.

This also explains the name 'philosophical logic': logics to analyze philosophical concepts. Sometimes, this is distinguished from 'philosophy

of logic': philosophical issues within logic. However, this distinction and the exact demarcation of what is 'philosophical logic' aren't fixed and opinions tend to differ over time.

Objectives In terms of content, the course aims to convey

1. many of the main formal logics found in the field of philosophical logic (many-valued logics, intuitionistic logic, relevance logic, counterfactuals, etc.)
2. an understanding of the philosophical concepts that these logics are built to analyze (truth, vagueness, paradoxes, etc.)

In terms of skills, the course aims to teach

1. the ability to build formal logics and mathematical models of (philosophical) phenomena. The tools that we learn to use in particular are algebraic semantics and state-based semantics
2. the ability to philosophically assess a proposed formalization.

Prerequisites A good introduction to logic, including the basics of formulating mathematical proofs.

Contents We start with a recap of classical logic, which we use, at the same time, to introduce more advanced content: (a) what, generally speaking, a logic is, and (b) what the two common ways are to provide a semantics for a logic (algebraic and state-based). Then we consider the main formal logics found in the field of philosophical logic (for a list, see the table of contents). The pattern of introducing them will always be, more or less, the following:

- First, we look at some important, mostly philosophical phenomenon (vagueness, inconsistent data, truth, paradoxes, etc.) where classical logic doesn't seem appropriate.
- Then we develop the formal logic as an alternative.
- Finally, we assess how well the logic serves as a model.

Omissions Given the limited time, we have to omit some topics that are usually considered to be part of philosophical logic. For example:

1. Quantification, existence (first-order logic). While quantification is an important and interesting topic, leaving it out has benefits: only looking at the propositional (and not first-order) part of a logic often makes its main features stand out more clearly. First-order logic (and possibly also the basics of second-order logic) usually is part of a good introductory course to logic.

2. Proof theory. We'll focus mostly on the semantic aspects of the logics covered.

3. Modal logic (necessity, possibility). There is a separate course on modal logic.

Layout As mentioned in the disclaimer, these notes are informal and partially still under construction. For example, there are margin notes to convey more casual comments that you'd rather find in a lecture but usually not in a book. Todo notes indicate, well, that something needs to be done. References are found at the end; as is an index to help looking up terms.

This is a margin note.

This is a todo note

Exercises There are four kinds of exercises (but the distinction isn't always sharp):

1. Exercises occurring in the main text. They encourage you to actively deal with the material rather than just passively absorbing it. They often are rather informally stated and don't need a fully formal proof, but should help the understanding.

The other three kinds are found at the end of a chapter and will usually be part of the homework assignment. (The terminology is partly that of Restall (2000).)

2. Practice exercises. They are meant to learn and reinforce concepts from the chapter. They usually involve untangling the definitions and some fairly straightforward reasoning steps.

3. Problem exercises. They usually require some creativity, e.g., an idea not mentioned in the chapter to fill in gaps in proofs or apply concepts of the chapter to other areas.

4. Philosophy exercises. Unlike the others, they don't have a clear formal solution, but rather pose a philosophical question which can have several answers. This could be a philosophical assessment of

a proposed formalization or a philosophical analysis of a concept. The answer should still use clear arguments, but need not be formal. Philosophy questions are a very different kind of 'difficult' than formal questions.

Study material

- G. Restall (2022). *Proofs and Models in Philosophical Logic*. Elements in Philosophy and Logic. Cambridge: Cambridge University Press
- G. Priest (2008). *An Introduction to Non-classical Logic. From If to Is*. 2nd ed. Cambridge: Cambridge University Press
- J. P. Burgess (2009). *Philosophical Logic*. Princeton Foundations of Contemporary Philosophy. Princeton: Princeton University Press
- G. Restall (2000). *An Introduction to Substructural Logics*. London: Routledge
- M. J. Dunn and G. M. Hardegree (2001). *Algebraic Methods in Philosophical Logic*. Oxford Logic Guides 41. Oxford: Oxford University Press
- T. Sider (2010). *Logic for Philosophy*. Oxford: Oxford University Press
- OpenLogicProject (2021)
- J. M. Font (2016). *Abstract Algebraic Logic: An Introductory Textbook*. Studies in Logic 60. London: College Publications
- The series "Handbook of Philosophical Logic" edited by D. Gabbay and F. Guenthner (Springer).

Acknowledgments Many thanks to both students and colleagues involved in the past and present editions of the course. In particular, thank you for very helpful comments: Sander Beckers, Elias Bronner, Isabella Cissell, Timo Diedering, Swapnil Ghosh, Søren Knudstorp, Frederik Lauridsen, Flip Lijnzaad, Alexander Lind, Johannes Marti, Armin Masala, Dean McHugh, Alyssa Reynaldi, Marie Schmidtlein, Zotescu Teodor-Ștefan, Jonathan Thul, Annica Vieser, and Wouter Vromen.

1 Prologue: paradoxes

Before starting in earnest, we'll look at a teaser for the course: paradoxes. They play a central role in philosophical logic by motivating many logics.

> A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science. (B. Russell 1905a, 484f.)

Liar paradox The most famous one arguably is the liar paradox. Consider the sentence

For much more, see, e.g., J. Beall, Glanzberg, et al. (2020).

This sentence is false. (1.1)

If sentence (1.1) is true, then what it says is the case, so (1.1) is false. If sentence (1.1) is false, then what it says is the case, so (1.1) is true. In short, (1.1) is true if and only if it is false.

You might have seen this too often to get excited. But appreciate how puzzling it was seeing it for the first time.

Here are two possible conclusions (Priest 2008, sec. 7.7): First, since a sentence is either true or false (aka bivalence), the sentence (1.1) is *both true and false*. (If (1.1) is true, it is also is false by the above; and if (1.1) is false, it also is true by the above.)

Second, if we, however, think that a sentence cannot be both true and false (aka law of non-contradiction), we must conclude that (1.1) is *neither true nor false*. However, then we face the *revenge paradox*

This sentence is either false or neither true nor false. (1.2)

If (1.2) is true, it is either false or neither, whence in any case not true. If (1.2) is not true, it can only be false or neither (assuming there are no further options), so what (1.2) says is the case, whence it is true. Again, since a sentence is either true or not true, (1.2) is both true and not true.

A pretty smart way of moving the paradox to the next level, isn't it?

This prompts us to investigate how a logic with the additional truth-values *neither true nor false* and/or *both true and false* looks like.

Card paradox (or liar cycles) One might dismiss the paradox by saying that a sentence referring to itself ('self-reference') is at best an artificial curiosity. But one can create the paradox also without explicit self-reference. Rather one only uses sentences saying about other sentences that they are true or false—and that's something we do (e.g., 'what they say is true').

Attributed to Philip Jourdain (O'Connor and Robertson 2005). Also called 'liar cycles', see J. Beall, Glanzberg, et al. (2020).

Sentence (1.4) is true. (1.3)

Sentence (1.3) is false. (1.4)

Or think of a card: the sentence on the front says that the sentence on the back is true, and the sentence on the back says that the one on the front is false.

Then, again, sentence (1.3) is true if and only if it is false: If (1.3) is true, then what it says is the case, so sentence (1.4) is true, so what it, in turn, says, is the case, i.e., sentence (1.3) is false. And if (1.3) is false, then what it says is not the case, so sentence (1.4) is not true, so what it says, is not the case, whence sentence (1.3) is not false, i.e., true.

Curry paradox This is a similar paradox, but it doesn't use negation or falsity (but self-reference), and yet arrives at the paradoxical conclusion that any sentence must be true: Let p be your favorite (false) sentence that you would like to be true. Now, consider

For more, see Shapiro and J. Beall (2021).

If sentence (1.5) is true, then p is true. (1.5)

We first show that sentence (1.5) is true: It's a conditional claim, so assume the antecedent and show the consequent. So we assume that sentence (1.5) is true, and show that p is true. Indeed, if (1.5) is true, then what it says is the case, so we have "if (1.5) is true, then p is true"; now since (1.5) is true by assumption, we also have (by modus ponens) that p is true, as needed.

A conditional sentence is of the form 'If A, then B' and A is called the antecedent and B is called the consequent.

Now, we get (with similar reasoning) that p is true: Since (1.5) is true (as shown), what it says is the case, so we have "if (1.5) is true, then p is true"; now since (1.5) is true (as shown), we also have (by modus ponens) that p is true, as needed.

Sorites paradox Let's move to another kind of paradox (at least on the face of it). Vague concepts like 'heap', 'bald', or 'tall' give rise to the sorites paradox.

In Greek, 'soros' means 'heap'.

- 1 grain of sand does not make a heap.
- If 1 grain doesn't make a heap, then 2 grains don't.
- If 2 grains don't make a heap, then 3 grains don't.

- ...
- If 999,999,999 grains don't make a heap, then 1 billion grains don't.

By applying modus ponens (if p and $p \to q$, then q) over and over again, these plausible premises imply that 1 billion grains of sand don't form a heap, which is clearly wrong.

For more, see Hyde and Raffman (2018).

This again prompts us to a careful analysis of the logical reasoning involved here.

There are many more (logical) paradoxes. And there also is work on determining the general underlying structure behind them: we get back to this in chapter 9. First and foremost, though, these paradoxes motivate us to (re)consider and possibly adapt our logic—which is what we'll do in this course.

For a list of paradoxes (more than you'd ever care to know), see `https://en.wikipedia.org/wiki/List_of_paradoxes`

2 Classical logic

We recap classical propositional logic. In doing so, we formulate as general templates what a logic is and two main ways of providing a semantics for a logic: state-based and algebraic. Classical logic fills in these templates in a particular way, and during the course we'll see many other ways these templates can be filled in.

Key concepts
- General concept of a logic
- Object-language vs. meta-language (vs. natural language)
- Syntax vs. semantics
- Propositional and Boolean language and their connectives
- Classical valuations, truth-tables
- Classical validity (tautology) and classical consequence
- Material conditional
- State-based semantics (template)
- Algebraic semantics, including truth-value semantics (template)
- Boolean algebra
- Equivalence of semantics

2.1 A logic template

What it means to 'define' a logic

Generally speaking, a logic aims to specify which reasoning steps are correct. One also speaks of argument (or inference) steps being valid. To do so, a logic specifies (or consists of) the following:

Template 2.1.

1. A formal language called the *object-language*. Its symbols (e.g., $\neg$ or $\wedge$) often are intended as formal analogues of natural language expressions (e.g., 'not' or 'and'). It is philosophically important to understand how close this relationship between the formal and the informal is.

2. A *proof system*, i.e., a set of rules describing when sentences can be 'derived' from other sentences (e.g., from the sentences φ and ψ you can formally derive the sentence $\varphi \wedge \psi$). When a sentence φ can be derived from a set of sentences Γ according to these formal rules, one writes $\Gamma \vdash \varphi$ and often speaks of *provability*. This is a matter

of purely formal symbol manipulation, whence one often calls the involved notions *syntactic*. However, the rules are often intended as formal analogues of typical informal reasoning patterns. Again, the relation between the formal and the informal is important.

3. A *formal semantics* which assigns a precise mathematical meaning to the sentences or the formulas (which, formally, are just meaningless symbols). Based on this, the semantics also provides a definition of when a set of sentences Γ implies another sentence φ, which one writes as $\Gamma \models \varphi$. Often this is done using a *meta-language*. For example, the statement "the sentence $\varphi \wedge \psi$ is true under a valuation if and only if both the sentences φ and ψ are true under that valuation" is a statement *about* the object-language.

This deliberately is formulated rather generally and vaguely, so the template can indeed accommodate the enormous range of existing logics. But we'll now consider a concrete example.

2.2 Classical propositional logic

Let's recap how classical propositional logic fills in this template.

The language of classical logic

Definition 2.2 (Propositional and Boolean language)**.** The symbols of the *propositional language* $\mathcal{L}_{\text{prop}} = \mathcal{L}(\vee, \wedge, \neg, \bot, \top, \to, \leftrightarrow)$ are infinitely many propositional variables (or atoms) $p_0, p_1, p_2, \ldots$ and the propositional connectives $\vee$ (disjunction), $\wedge$ (conjunction), $\neg$ (negation), $\bot$ (bottom/falsum), $\top$ (top/verum) $\to$ (conditional), $\leftrightarrow$ (biconditional), and the punctuation marks (and). The $\mathcal{L}_{\text{prop}}$-sentences (aka propositional formulas) are generated from the propositional variables by the rule

> If φ and ψ are sentences, then so are $(\varphi \vee \psi), (\varphi \wedge \psi), \neg\varphi, \bot, \top, (\varphi \to \psi), (\varphi \leftrightarrow \psi)$.

The *Boolean language* $\mathcal{L}_{\text{bool}} = \mathcal{L}(\vee, \wedge, \neg, 0, 1)$ is defined in the same way but doesn't use the connectives $\to$ and $\leftrightarrow$.

Notational conventions

The usual conventions apply: We write $\mathsf{P} := \{p_0, p_1, \ldots\}$ for the set of propositional atoms. We use lower-case Greek letters $\varphi, \psi, \chi, \ldots$ as variables ranging over formulas. We use $p, q, r, \ldots$ as variables ranging over propositional atoms. We use upper-case Greek letters $\Gamma, \Delta, \Sigma, \ldots$ as variables ranging over sets of formulas (the empty set is written $\emptyset$). We often omit parentheses when there is no danger of confusion (e.g., write $(\varphi \wedge \psi)$ simply as $\varphi \wedge \psi$). We also write $\mathcal{L}_{\text{prop}}$ for the set of $\mathcal{L}_{\text{prop}}$-sentences, similarly for $\mathcal{L}_{\text{bool}}$. We write 'iff' for 'if and only if'.

There are many proof systems for classical logic: Hilbert calculi, natural deduction systems, sequent calculi, tableaux systems, etc. You probably know some from your introduction to classical logic. Since we won't need them here, we don't specify them and move straight to semantics.

The proof system(s) of classical logic

Definition 2.3 (Valuation)**.** A *(classical) valuation* is a function $\nu : \mathsf{P} \to \{0, 1\}$ from the set of propositional atoms to the set of truth-values 0 (false) and 1 (true). We recursively extend ν from the propositional atoms to all formulas by:

The semantics of classical logic

- $\nu(\varphi \vee \psi) = 1$ if $\nu(\varphi) = 1$ or $\nu(\psi) = 1$; and $= 0$ otherwise.
- $\nu(\varphi \wedge \psi) = 1$ if $\nu(\varphi) = 1$ and $\nu(\psi) = 1$; and $= 0$ otherwise.
- $\nu(\neg\varphi) = 1$ if $\nu(\varphi) = 0$; and $= 0$ otherwise.
- $\nu(\bot) = 0$.
- $\nu(\top) = 1$.
- $\nu(\varphi \to \psi) = 1$ if $\nu(\varphi) = 0$ or $\nu(\psi) = 1$; and $= 0$ otherwise.
- $\nu(\varphi \leftrightarrow \psi) = 1$ if $\nu(\varphi) = \nu(\psi)$; and $= 0$ otherwise.

If $\nu(\varphi) = 1$ (resp. $= 0$), we say the sentence φ is true (resp. false) according to the valuation ν (or ν makes true φ) and also write $\nu \vDash \varphi$. We write $\nu \nvDash \varphi$ if it is not the case that $\nu \vDash \varphi$.

A set of sentences Γ *implies* a sentence φ (or φ is a *consequence* of Γ), written $\Gamma \vDash \varphi$, iff for any valuation ν, if ν makes true every sentence in Γ, then ν makes true φ. If $\emptyset \vDash \varphi$, i.e., every valuation makes true φ, we call φ a *logical truth* (or *tautology*). (If φ is true under some valuation, we say it is *satisfiable*.) We write $\Gamma \nvDash \varphi$ if it is not the case that $\Gamma \vDash \varphi$. Two sentences φ and ψ are logically equivalent iff $\varphi \vDash \psi$ and $\psi \vDash \varphi$.

Later, when we also deal with logics different from classical logic, we write $\vDash_{\mathsf{CL}}$ to stress that we're considering the consequence relation of classical logic.

Remark 2.4. A convenient way to represent and calculate valuations is with truth-tables: see figure 2.1. It shows that the sentences $\varphi \to \psi$ and $\neg\varphi \vee \psi$ are logically equivalent. If this is the case in a logic, one says $\to$ is the *material conditional*. Consequently, we can also leave out the connective $\to$ (and similarly for $\leftrightarrow$) and only treat it as being *defined* as (or an abbreviation of) $\varphi \to \psi := \neg\varphi \vee \psi$. This is why classical logic is often only formulated in the Boolean language.

There are further redundancies: e.g., $\varphi \wedge \psi$ is equivalent to $\neg(\neg\varphi \vee \neg\psi)$.

φ	ψ	$\neg\varphi$	$\varphi \to \psi$	$\neg\varphi \vee \psi$
0	0	1	1	1
0	1	1	1	1
1	0	0	0	0
1	1	0	1	1

Figure 2.1: A truth-table showing that $\varphi \to \psi$ and $\neg\varphi \vee \psi$ are equivalent.

Exercise 2.5. Cover the 0's and 1's in figure 2.1 and make sure you can fill them back in correctly yourself.

Remark 2.6. Note that we can give the propositional language a different proof system and semantics (we'll later see, e.g., that of intuitionistic logic). Then we obtain a different logic—albeit one over the same language—and it might be that $\varphi \to \psi$ is *not* equivalent to $\neg\varphi \vee \psi$ (as in intuitionistic logic). Then the Boolean language is a *proper* fragment of the propositional language.

Exercise 2.7. Go back to the logic template from section 2.1 and make sure you understand how exactly classical propositional logic fills in the three ingredients 'language', 'proof system', and 'semantics' (though we didn't say much about proof systems).

2.3 Two kinds of semantics

The semantics of classical logic in terms of valuations can be viewed from two perspectives: state-based and algebraic. In the next section, we show that they are equivalent—which is one of the deepest results about classical propositional logic.

A main theme of the course will be to explore the two perspectives in tandem: Each provides an important and distinct perspective on (developing a semantics for) a given logic. And if they agree, one can reap their respective advantages.

The slogans of the two perspectives—that we will explain in the remainder of this chapter—are:

- *State-based semantics*: first states, then propositions! States are the basic concept: possibilities at which sentences or propositions (i.e., meanings of sentences) are true. Propositions are derived as truth-sets: the set of states at which they are true.

- *Algebraic semantics*: first propositions, then states! Propositions are the basic concept: the meanings of sentences, which have algebraic structure corresponding to the sentential connectives. States are derived as the possible models of these sentences/propositions.
- *Stone duality*: Going, on the one hand, from states to propositions and, on the other hand, from propositions to states renders these two semantics equivalent.

The formal statement is that, when formulated properly as categories, the two are dual—hence the name. See Gehrke (2009) for an introduction to duality for a general audience, describing it as: "the two aspects are not separate and unrelated but different aspects of the same thing" (p. 10).

2.3.1 State-based semantics

A valuation $v : \mathsf{P} \to \{0, 1\}$ can be regarded as a 'possible world': We think of the elements $p_0, p_1, \ldots$ of P as describing the possible atomic facts (or state of affairs) of how the world can be; and the valuation v says which of them obtain and which don't in the possible world that v is describing. Note that these possible worlds hence are complete (every p_i is either true or false, none is left unspecified) and consistent (no p_i is both true and false).

There is much discussion in philosophy about what possible worlds are (Menzel 2021). Taking them as complete and consistent sets of sentences (as we essentially do) is a position known as combinatorialism or linguistic ersatzism. Here we won't discuss the metaphysics of possible worlds, and to indicate that we'll more neutrally speak of *(classical) states*. Later on, we'll encounter other, 'non-classical' kinds of states, e.g., some that aren't complete or consistent. Thus, states can be many things: from the possible states of knowledge of a reasoner to the states of a dynamical system.

In fact, we can decouple what a state *is* from what it *does*, namely its ability to make true or false (atomic) formulas. This idea yields the template for state-based semantics:

This is just sketching a general idea, not a precise definition. It now sounds awfully abstract, but we'll soon fill it with life. The only point to understand now is that there is a general pattern that we'll fill in concretely in various ways during the course.

Template 2.8 (State-based semantics)**.** 1. A *model* of a state-based semantics is a triple $\mathsf{M} = (\mathsf{S}, \mathsf{R}, \mathrm{I})$ where

   - S is a set of states (also called the *state space*), usually required to be nonempty,
   - R is a collection of *relations* (including functions) on S,
   - and I, called *interpretation*, describes, for each state $s \in \mathsf{S}$, which atomic formulas it makes true (and which false),

   and further conditions may be imposed on M.

2. While I describes truth and falsity of atomic sentences, the relations in R (if there are any) are used to interpret other connectives of the language. Thus, we can write, for a state $s \in S$ and a sentence φ of the language, $M, s \models \varphi$ to say that φ is true at state s (or s makes true φ) in model M according to the present state-based semantics.

3. The set $[\![\varphi]\!] := \{s \in S : M, s \models \varphi\}$ of states making true φ is also called the *truthset* of φ. It can be seen as the meaning of sentence φ (or the proposition that φ expresses).

4. We define $\Gamma \models \varphi$ if for all models $M = (S, R, I)$ and states $s \in S$, if s makes true every sentence in Γ, then s also makes true φ.

5. The part (S, R) that forgets the interpretation is called a *frame* of the state-based semantics.

Example 2.9. In the case of classical logic, a *classical state-based model* $M = (S, I)$ is just a set S together with a function $I : S \times \mathsf{P} \to \{0, 1\}$ such that, for each $s \in S$, $v_s := I(s, \cdot) : \mathsf{P} \to \{0, 1\}$ is a classical valuation. So we don't make use of any relations, and I determines, for each state, which atomic sentences it makes true. The relation $M, s \models \varphi$ is then defined as in definition 2.3: $M, s \models \neg\varphi$ iff $M, s \not\models \varphi$; $M, s \models \varphi \wedge \psi$ iff $M, s \models \varphi$ and $M, s \models \psi$; etc. ⌟

Of course, for classical logic this terminology is needlessly complicated. But the point is that different choices for the parameters in the template will give different semantics for various logics that we'll encounter in the course. Maybe the most well-known example for such a state-based semantics is modal logic (Blackburn et al. 2001):

Example 2.10. The language of modal logic $\mathcal{L}(\vee, \wedge, \neg, 0, 1, \Box)$ extends the Boolean language $\mathcal{L}_{\text{bool}}$ with an additional unary connective $\Box$ (i.e., if φ is a sentence, also $\Box\varphi$ is). A *basic model* of a state-based semantics for modal logic (i.e., Kripke semantics) is a structure $M = (S, R, I)$ where S is a set, $R \subseteq S \times S$ a binary relation, and $I : S \times \mathsf{P} \to \{0, 1\}$ a function such that each $I(s, \cdot)$ is a classical valuation. And the relation R is used to interpret the connective $\Box$:

- $M, s \vDash p$ iff $I(s, p) = 1$
- $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$ (similarly for $\vee$)
- $M, s \vDash \neg\varphi$ iff $M, s \nvDash \varphi$
- $M, s \vDash \bot$ never holds, and $M, s \vDash \top$ always holds

Only the last bullet point is new. Its intuition is explained below.

- $M, s \vDash \Box\varphi$ iff for all $s' \in S$, if sRs', then $M, s' \vDash \varphi$.

The intuition is to interpret sRs' as saying that state s' is *accessible* from state s. Then $M, s \vDash \Box\varphi$ means that φ is true in all states accessible from s. This is used to interpret necessity (truth in all relevant/accessible possible worlds) or knowledge (truth in all situations that are possible according to what I currently know). However, as there is a dedicated course on modal logic, we won't deal with modal logic here. ⌟

We'll see many more examples of state-based semantics below (for databases, truth-making, counterfactuals, etc.)

2.3.2 Algebraic semantics

The other perspective on the semantics of classical logic is algebraic, which we explore in this subsection.

What's 'algebraic' about this? Algebra is the study of rules for combining objects or symbols. In school, this means combining symbols like $ax^2 + bx + c$ and studying when they equal another, like 0. Here it will be about combining propositions or truth-values.

Propositions The task of a semantics is to assign meaning to the sentences of the language. The meaning of a sentence is also called a *proposition*. As logicians, we don't really care what exactly propositions *are* but we do care about how they *behave* (or relate to each other). After all, in logic we care about which sentences are tautologies (true under any interpretation/assignment of meaning) and consequences of others (truth-preserving under any interpretation/assignment of meaning).

For example, given a proposition a (say the meaning of the sentence 'The sun shines') and a proposition b (say the meaning of the sentence 'It is warm') there also is a proposition $a \wedge b$ (which is the meaning of the sentence 'The sun shines and it is warm'). This proposition should be identical to the proposition $b \wedge a$ (which is the meaning of the yet different sentence 'It is warm and the sun shines'). In symbols, $a \wedge b = b \wedge a$. Similarly, we also should have propositions $a \vee b$ (which is the meaning of the sentence 'The sun shines or it is warm'), $\neg a$ (which is the meaning of the sentence 'The sun does not shine'), 0 (which is the meaning of a contradiction like 'The sun shines and does not shine'), and 1 (which is the meaning of a tautology like 'The sun shines or does not shine'). And these, too, should satisfy the expected equations (reminiscent of equivalences in classical logic), like $a \wedge 0 = 0$ or $a \vee a = a$.

Thus, we have a set A of propositions (whatever objects they are), and we have two binary functions $\wedge, \vee : A \times A \to A$, one unary function $\neg : A \to A$, and two constants (or 0-ary functions) $0, 1 \in A$. And they should satisfy various equations like the ones above. What equations we

x	y	$f_\vee(x,y)$
0	0	0
0	1	1
1	0	1
1	1	1

Figure 2.2: The truth-function ($f_\vee$) corresponding to disjunction ($\vee$).

exactly want to require is not a trivial matter, but those that turn out to work are listed in definition 2.11 below. The structures $(A, \vee, \wedge, \neg, 0, 1)$ that satisfy these required equations are known as *Boolean algebras*.

The fact that we don't just consider a single such Boolean algebra but the whole class of them reflects the logicians' view on propositions: It doesn't matter what propositions are, i.e., what the underlying set A is. This may be because we don't care to determine them or because there are various possible meaning assignments depending on context, etc. What matters is that the propositions carry the right structure, i.e., the functions $\vee, \wedge, \neg, 0, 1$ satisfying the required equations.

Truth-value semantics There also is another way to arrive at the concept of a Boolean algebra. (At the end of this paragraph we relate it to the preceding way.) To see this, let's reformulate the valuation semantics a little.

We have a set of truth-values $T = \{0, 1\}$. The sentence connectives $\vee, \wedge, \neg$ are interpreted as functions on T given by truth-tables: For example, $\vee$ is interpreted as the function $f_\vee : T \times T \to T$ defined in figure 2.2. These functions are also called *truth-functions*. If we think of T as being ordered as usual (i.e., $0 < 1$), we can also write $f_\vee(x, y) = \max(x, y)$ (i.e., $f_\vee(x, y)$ is the maximum of the two numbers x and y). Similarly, $\wedge$ is interpreted by $f_\wedge : T \times T \to T$ given by $f_\wedge(x, y) = \min(x, y)$, and $\neg$ is interpreted by $f_\neg : T \to T$ given by $f_\neg(x) = 1 - x$. The propositional constants $\bot$ and $\top$ are interpreted as elements from T: $f_\bot := 0$ and $f_\top := 1$.

Exercise: verify this!

So the valuation semantics interprets the connectives as truth-functions on the set of truth-values T. We can summarize this as the structure $(T, f_\vee, f_\wedge, f_\neg, f_\bot, f_\top)$. The notation is usually simplified: First, we write $2 = \{0, 1\}$, since, in set theory, a natural number is defined to be the set of its predecessors. Second, we also write $\vee$ for $f_\vee$, and $\neg$ for $f_\neg$, etc., since context should make clear whether we mean the connective or the truth-function. So we write $(2, \vee, \wedge, \neg, 0, 1)$, which is abbreviated as **2**.

Crucially, the truth-functions operating on the set of truth-values satisfy

various equations. For example, for all possible values of x and y in 2, we have

$$x \wedge x = x \qquad x \vee (x \wedge y) = x \qquad x \wedge \neg x = 0. \tag{2.1}$$

Exercise: verify this!

This invites a question typical in algebra: How crucial is it to work with the concrete choice of the structure **2**? Can we also work with more general structures that resemble **2** only on its essential structure—in the sense of satisfying its characteristic equations? There is no universal way of finding such an appropriate definition of a general structure, but here it is that of a Boolean algebra. The idea is as follows:

Maybe you've heard of the following examples: Distilling the general structure of a field *from, say, the rational numbers* $\mathbb{Q}$*, or of a* vector space *from* $\mathbb{R}^n$*, or of a* group *from the symmetric group* S_n.

- We generalize the underlying set from $2 = \{0, 1\}$ (the set of classical truth-values) to any non-empty set A (and think of its elements as 'generalized' truth-values).
- We still require the functions of the right *signature*: two 2-ary functions $\vee, \wedge : A \times A \to A$, one 1-ary function $\neg : A \to A$, and two 0-ary functions $0, 1 \in A$. (Sometimes one writes the signature as $(2, 2, 1, 0, 0)$ to indicate how many functions of which arity are required.)
- And, finally, we also require these functions to satisfy the equalities that are characteristic for the truth-tables, like those in 2.1. The precise definition is stated below (definition 2.11).

You might come up with different answers!

How the two motivations for Boolean algebras relate? One answer is: In a sense, the most general truth-values for sentences are the propositions expressed by them; they, too, can be ordered by 'how true' they are (or how close they are to the tautology 1) and form a Boolean algebra.

Boolean algebras Now, that we've motivated Boolean algebras, let's introduce them formally in the style of a math textbook.

Don't worry about the details of this definition. You can always come back should you need it.

Definition 2.11. A *Boolean algebra* is a structure $(A, \vee, \wedge, \neg, 0, 1)$ where A is a nonempty set, $\vee, \wedge : A \times A \to A$ are binary functions, $\neg : A \to A$ is a unary function, and $0, 1 \in A$ are such that

Some write x' *instead of* $\neg x$.
A structure $(A, \vee, \wedge)$ *satisfying these axioms is known as a* lattice. *(Idempotence is already implied by absorption.)*

1. Lattice axioms:
   a) Commutativity: $x \vee y = y \vee x$ and $x \wedge y = y \wedge x$
   b) Associativity: $x \vee (y \vee z) = (x \vee y) \vee z$ and $x \wedge (y \wedge z) = (x \wedge y) \wedge z$
   c) Idempotence: $x \vee x = x$ and $x \wedge x = x$

d) Absorption: $x \vee (x \wedge y) = x$ and $x \wedge (x \vee y) = x$.

2. Distributivity: $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$ and $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$.

3. Least and greatest elements: $x \wedge 0 = 0$ and $x \vee 1 = 1$

4. Complements: $x \wedge \neg x = 0$ and $x \vee \neg x = 1$.

Write $\wedge$ as $\times$ (multiplication) and $\vee$ as $+$ (addition) and the second equality is the distributivity you know from numbers: $x(y + z) = xy + xz$.

We often just write A to mean the Boolean algebra $(A, \vee, \wedge, \neg, 0, 1)$. We define $x \leqslant y$ as $x \wedge y = x$ (or, equivalently, as $x \vee y = y$), which defines a partial order on A. Observe that axiom 3 then says that 0 is the least element (i.e., $\forall x \in A : 0 \leqslant x$) and 1 is the greatest element (i.e., $\forall x \in A : x \leqslant 1$).

The definition of a partial order *is in the appendix.*

Exercise 2.12. Show that $x \wedge y = x$ and $x \vee y = y$ are indeed equivalent (hint: use absorption), and show that $\leqslant$ indeed defines a partial order.

For context: Boolean algebras are an instance of the general idea of an *algebraic structure* $(A, f_1, \ldots, f_n)$: as set A together with some functions $f_1, \ldots, f_n$ of finite arity defined on A (usually required to satisfy various equations like the axioms above). These algebraic structures are studied in the field universal algebra (Burris and Sankappanavar 1981; Grätzer 2008). You'll learn more about this—and Boolean algebras in particular—in the course 'Mathematical Structures in Logic'.

Some of the most important examples are the following:

Exercise 2.13.

1. Show that **2** is a Boolean algebra, i.e., that its operations satisfy the above axioms. This is depicted on the left of figure 2.3.

2. Show that, for any set X, its powerset $\mathcal{P}(X)$ together with the set-theoretic operations of union ($\cup$), intersection ($\cap$), and complement (^{c}), and the constants $0 = \emptyset$ and $1 = X$ forms a Boolean algebra (and that the order $\leqslant$ is the subset relation $\subseteq$).

3. Convince yourself of the following: If $X = \emptyset$ is the empty set, then $\mathcal{P}(X) = \{\emptyset\}$ is the 'trivial' Boolean algebra with just one element (hence $0 = 1$). If $X = \{a\}$ is a set with just one element, then $\mathcal{P}(X) = \{\emptyset, \{a\}\}$ is essentially the Boolean algebra **2** with $0 = \emptyset$ and $\{a\} = 1$. If $X = \{a, b\}$ is a set with two elements, then $\mathcal{P}(X) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$ is the Boolean algebra depicted on the right of figure 2.3.

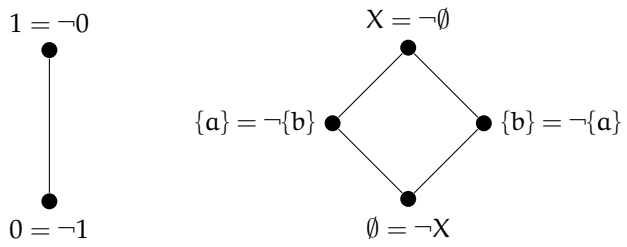

Figure 2.3: Two examples of Boolean algebras: **2** on the left and $\mathcal{P}(X)$ with $X = \{a, b\}$ on the right.

Exercise 2.14. If you'd like to get acquainted more with the axioms, a good and challenging exercise is to prove that the axioms imply the following well-known identities:

1. Double negation elimination: $\neg\neg x = x$.
2. De Morgan: $\neg(x \wedge y) = \neg x \vee \neg y$ and $\neg(x \vee y) = \neg x \wedge \neg y$.

Hint: First show that $x \vee y = 1$ and $x \wedge y = 0$ implies $\neg x = y$.

Algebraic semantics formally With the concept of a Boolean algebra, we can finally state the algebraic semantics for classical logic. In fact, we first state the general abstract template of an algebraic semantics for a logic, and then we fill it in for the case of classical logic.

Template 2.15 (Algebraic semantics)**.** 1. Single out a class $\mathcal{C}$ of algebraic structures with operations $\vee, \wedge, \neg, 0, 1$ to interpret the Boolean language (and add further operations if we consider more connectives).

2. For $A \in \mathcal{C}$, an A-*valuation* is a function $v : \mathsf{P} \to A$, recursively extended to formulas by

$$v(\varphi \wedge \psi) := v(\varphi) \wedge v(\psi) \qquad v(\bot) := 0$$
$$v(\varphi \vee \psi) := v(\varphi) \vee v(\psi) \qquad v(\top) := 1$$
$$v(\neg\varphi) := \neg v(\varphi).$$

By the notational conventions, $\wedge, \vee, \ldots$ are connectives on the left of $=$ and algebraic operations on the right.

3. For a possibly empty set of sentences Γ and a sentence φ, define $\Gamma \vDash \varphi$ by

   For all $A \in \mathcal{C}$ and A-valuations v, if $v(\psi) = 1$ for all $\psi \in \Gamma$, then $v(\varphi) = 1$.

There also are generalizations of this definition. But for now, this is enough.

(See Blok and Pigozzi (1989), Font (2016), or Jansana (2022) for this and a more general formulation.)

Exercise 2.16. To obtain the usual semantics of classical logic using valuations just set $\mathcal{C} := \{\mathbf{2}\}$, i.e., the class $\mathcal{C}$ of algebraic structures just consists of the single two-element Boolean algebra **2**. Verify that the semantics obtained from the template with this choice of $\mathcal{C}$ really is just the valuation semantics from definition 2.3.

If $\mathcal{C}$ consists of a single (typically finite) algebraic structure, we may speak of a *truth-value semantics*: it gives a semantics to the logic by means of interpreting its connectives as (truth-) functions over the (finite) set of truth-values. Thus, we can see this as a special kind of algebraic semantics. In the next chapter, we study this kind of semantics when we have three (or four) rather than just two truth-values.

An advantage of this approach is that it defines the semantics with respect to a single algebra of truth-values. However, as motivated above, we also want to consider (a large part of) the class of all Boolean algebras: Either (1) as representing the possible choices of propositions (hence avoiding commitment to what objects they are), or (2) as representing only the structural aspects of truth-values (hence avoiding the explicit choice of 0 and 1). Thus, the semantics picks out the structural aspects shared by all members of the class of algebras rather than the idiosyncrasies of a single member.

By analogy, an algebraist studying, say, groups doesn't just want to study a single group like the S_n but also wants to understand the structure behind larger classes of groups.

Example 2.17. In the case of classical logic, we have already seen a natural candidate for a larger class of algebras: namely choosing $\mathcal{C}$ as the class BA of all Boolean algebras. ⌟

Amazingly, these two semantics—the truth-value semantics choosing $\mathcal{C} := \{\mathbf{2}\}$ and the algebraic semantics choosing $\mathcal{C} := \mathsf{BA}$—are equivalent. This is the main result of this chapter (which we state in the next section). So even though Boolean algebras can be very complicated and very large, they don't add anything from the logical point of view to the very simple and small Boolean algbera **2**.

Make sure you really appreciate how amazing this is—then it's more fun to look at the proof.

Example 2.18. To get some intuition for why the two versions of algebraic semantics for classical logic coincide, let's consider the Boolean algebra $\mathbf{4} := \mathcal{P}(\{a, b\})$ from the right of figure 2.3. Looking at 'the logic of the powerset of a two-element set' means looking at which sentences evaluate to 1 under any **4**-valuation v. And we find the typical examples known from classical logic. For example: $v(p \vee \neg p) = 1$ because, writing $x :=$

$v(p) \in \mathbf{4}$, we have $v(p \vee \neg p) = v(p) \vee \neg v(p) = x \vee \neg x = 1_{\mathbf{4}}$ (by the complements axiom). ⌟

2.4 Equivalence of the algebraic and state-based semantics for classical logic

In this section, we show that the two approaches to semantics—state-based and algebraic—are in fact equivalent. Results of this form go under the name of *Stone duality*. We won't properly go into this mathematical theory; we just take the main ideas that we need here.

For this, see the course "Mathematical Structures in Logic". The standard math reference is Johnstone (1982).

More precisely, we show that the four semantics for classical logic that we've seen so far are all equivalent—even though they represent rather different perspectives on semantics.

Theorem 2.19. *Let φ be a sentences in the propositional language $\mathcal{L}_{\text{prop}}$ and Γ be a set of such sentences. Then the following are equivalent:*

That's one of the deepest results about classical propositional logic.

1. $\Gamma \models \varphi$ *according to the usual valuation semantics of definition 2.3.*
2. $\Gamma \models \varphi$ *according to the state-based semantics of example 2.9.*
3. $\Gamma \models \varphi$ *according to the truth-value semantics of exercise 2.16.*
4. $\Gamma \models \varphi$ *according to the algebraic semantics of exercise 2.17.*

Regarding the proof, the more straightforward parts of the theorem are the following exercises.

Exercise 2.20. Show (1)⇔(2).

Exercise 2.21. Show (1)⇔(3). (This was exercise 2.16.)

Exercise 2.22. Show (4)⇒(3).

The main part of the theorem is the implication (3)⇒(4). We'll establish this step by step in the exercises (section 2.5), where we use some ideas from Stone duality: namely, how to go back and forth between the algebraic and the state-based semantics.

We end this section by summarizing—and adding to—the respective advantages of state-based semantics and algebraic semantics. And we say a word on historic significance.

- The standard valuation semantics can be seen both as a state-based semantics (valuations as possible worlds) and as an algebraic semantics (as truth-value semantics using only the Boolean algebra **2**).

- State-based semantics (and also truth-value semantics) have the advantage of being rather intuitive. It has a *local* perspective on propositions: when they are true at a state (or at a valuation).
- Algebraic semantics has the advantage of being very structural. It has a *global* perspective on propositions: how they can be combined with other propositions to produce new ones.

The advantage of theorems about the coincidence of the state-based semantics and the algebraic semantics is that their respective advantages can be combined: the intuitiveness with the structure; the local with the global; etc. A typical example is to prove soundness and completeness: that $\Gamma \models \varphi$ holds iff φ is derivable from Γ in a given proof system. This is rather elegantly done in the algebraic semantics for $\models$ since, due to its focus on laws, it is closer to the rules of the proof system. Using the coincidence, this then automatically carries over to the state-based semantics for $\models$ with its intuitive terminology of states making true sentences.

We'll see this in chapter 4.

In the history of technology, Boolean algebras became important when Claude E. Shannon wrote his master's thesis about how they can be used in electrical engineering (Shannon 1940). The idea is that combining switches into electrical circuits resembles forming complex sentences from atomic ones: putting two switches p and q in series results in a circuit that is closed (0, current can flow) iff the two switches p and q are closed, and hence acts like $p \vee q$. So logic could be used for designing and analyzing such electric circuits—which is central to all digital computers. From Shannon's introduction:

Yes, the one from information theory.

One might also swap it: represent closed by 1 *and use* $\wedge$.

> Any circuit is represented by a set of equations, the terms of the equations representing the various relays and switches of the circuit. A calculus is developed for manipulating these equations by simple mathematical processes, most of which are similar to ordinary algebraic algorisms [*sic*]. This calculus is shown to be exactly analogous to the Calculus of Propositions used in the symbolic study of logic (Shannon 1940, p. 2).

Shannon (1940, p. 8) mentions the two interpretation of the symbols of Boolean algebras that we show to be equivalent here: either ranging over the elements of any Boolean algbera ("the algebra of classes") or over the elements of **2** ("the calculus of propositions"). With his work, he thus adds a third interpretation in terms of electric circuits.

2.5 Exercises

These exercises establish the main part of theorem 2.19. As mentioned, they contain the main ideas of what's known as Stone duality theory. The last exercise asks to reflect on what these ideas mean for the philosophers' favorite: possible worlds semantics.

We want to establish the implication (3)⇒(4). Let's first sketch the idea how to do this (and afterward explain why one might have it). We assume the following about our Γ and φ:

(a) For any **2**-valuation ν, if $\nu(\psi) = 1$ for all $\psi \in \Gamma$, then $\nu(\varphi) = 1$.

Given a Boolean algebra A and an A-valuation ν with $\nu(\psi) = 1$ (so 1 is the top element of A) for all $\psi \in \Gamma$, we need to show that also $\nu(\varphi) = 1$. We do this as follows:

1. Embed A into a power $\mathbf{2}^X$ of the Boolean algebra **2**.
2. If we evaluate a sentence $\psi \in \Gamma$ according to ν not just in A but in the bigger context of $\mathbf{2}^X$, it still gets value 1, so each component evaluates to 1 in **2**.
3. Then the assumption (a) implies, on each component, that also φ evaluates to 1 there. So φ evaluates to 1 in the bigger context of $\mathbf{2}^X$, and hence also in the smaller (original) context A, as needed.

This is the set-theoretic notion of 'exponentiation': the elements of $\mathbf{2}^X$ are of the form $(a_i : i \in X)$ with $a_i \in \mathbf{2}$. The connection to usual exponentiation is this: If X is finite with n-many elements, then the set $\mathbf{2}^X$ has 2^n-many elements.

How would one come up with this idea? Sure, given assumption (a) one would expect to do something with the Boolean algebra **2**, and once (1) is in place, steps (2) and (3) are quite natural. But why would one come up with (1) exactly? Here is the philosophical idea:

- As motivated, think of the elements of A as the possible propositions under consideration. Propositions are meaning of sentences, and an essential feature of the meaning of a sentence is to tell us, given a possible world (or model), whether the sentence is true or false in that world. So let X be the set of 'possible worlds'. The fact that the 'essence' of a proposition $a \in A$ is given by the function $X \to 2$ it determines is the idea behind the embedding $A \to \mathbf{2}^X$.
- That's the key idea of Stone duality: If we have a notion of state, we can form an algebra of propositions by taking the truthsets of sentences. But, crucially, we can also go the other way: We start with an algebra A of propositions, and develop a notion of state

What does the existence of this embedding say on the generalized truth-value motivation of Boolean algebras? A general truth-value $a \in A$ can be represented as a classical truth-value relativized *to a set of parameters X.*

(or 'model' or 'world'; in the case of classical logic, these are the so-called ultrafilters on A). Then we represent each proposition $a \in A$ by a truthset, or, equivalently, by a function $f : A \to 2$ (which says whether or not a proposition b is in this truthset).

Making this idea formal and filling in the gaps is the task of the exercises. Let's start with this idea of a power Boolean algebra:

Exercise 2.a (Practice)**.** Write $\mathbf{2} = (2, \vee_2, \wedge_2, \neg_2, 0_2, 1_2)$ thinking $2 = \{0, 1\}$. Let X be a set. Define the Boolean algebra $\mathbf{2}^X := (A, \vee, \wedge, \neg, 0, 1)$ by

- $A := \prod_X 2$, i.e., A is the set of sequences of the form $(a_i : i \in X)$ with $a_i \in 2$
- $(a_i : i \in X) \vee (b_i : i \in X) := (a_i \vee_2 b_i : i \in X)$
- $(a_i : i \in X) \wedge (b_i : i \in X) := (a_i \wedge_2 b_i : i \in X)$
- $\neg(a_i : i \in X) := (\neg_2 a_i : i \in X)$
- $0 := (0_2 : i \in X)$
- $1 := (1_2 : i \in X)$.

Formally, such a sequence is the function $X \to 2$ mapping $i \mapsto a_i$

One also says: the operations are defined elementwise.

Show that $\mathbf{2}^X$ is a Boolean algebra.

Next, we specify what an 'embedding' should be. To do so, we first need to take a step back: Generally speaking, when we consider sets with additional structure (like Boolean algebras, but also like groups, vector spaces, etc.), we're usually not interested in any function between the sets, but in those that preserve the relevant structure. Such functions then get the fancy name 'homomorphism'. Embeddings then are homomorphisms that also are injective as a function. Formally, for Boolean algebras, this reads as follows.

A function $f : X \to Y$ is injective *if, for all $x, y \in X$ with $x \neq y$ also $f(x) \neq f(y)$. So f doesn't identify distinct elements.*

Definition 2.23 (BA-homomorphism)**.** Let $A = (A, \vee_A, \wedge_A, \neg_A, 0_A, 1_A)$ and $B = (B, \vee_B, \wedge_B, \neg_B, 0_B, 1_B)$ be Boolean algebras. A *Boolean algebra homomorphism* (or just *BA-homomorphism*) from A to B is a function $f : A \to B$ such that, for all $a, b \in A$,

1. $f(a \vee_A b) = f(a) \vee_B f(b)$
2. $f(a \wedge_A b) = f(a) \wedge_B f(b)$
3. $f(\neg_A a) = \neg_B f(a)$
4. $f(0_A) = 0_B$

Meditate on how these conditions specify that the function f is preserving the relevant structure

5. $f(1_A) = 1_B$.

A BA-homomorphism $f : A \to B$ is a *BA-embedding* if f is injective.

Note that, since we write A both for the set A and for the Boolean algebra $(A, \vee, \wedge, \neg, 0, 1)$, we can also write $f : A \to B$ both for the function between the sets A and B but also for the BA-homomorphism. Context makes clear which is meant.

To get familiar with this concept, show the following.

Exercise 2.b (Practice)**.** Consider again the Boolean algebra $\mathbf{2}^X$ from the previous exercise. For $j \in X$, define the function

$$\begin{aligned} p_j : \mathbf{2}^X &\to \mathbf{2} \\ (a_i : i \in X) &\mapsto a_j. \end{aligned}$$

Show that p_j is a BA-homomorphism. Such functions are known as *projections*.

Now we're ready for the crucial part of the idea (step 1): the embedding $A \to \mathbf{2}^X$. The elements of X, i.e., intuitively the states (or models)—are defined as so-called ultrafilters on A We first state the formal definition, then motivate it.

This is a key concept from Stone duality and mathematical logic more broadly.

Definition 2.24 (Ultrafilter)**.** Let $A = (A, \vee, \wedge, \neg, 0, 1)$ be a Boolean algebra. A subset $U \subseteq A$ is an *ultrafilter* on A iff

1. Upset: For $a, b \in A$ with $a \leqslant b$, if $a \in U$, then also $b \in U$.
2. Closure: For $a, b \in A$, if both $a \in U$ and $b \in U$, also $a \wedge b \in U$.
3. Nonempty: $1 \in U$.
4. Proper: $0 \notin U$.
5. Ultra: For each $a \in A$, either $a \in U$ or $\neg a \in U$.

A subset $F \subseteq A$ satisfying conditions 1–3 is called a *filter*.

A typical example of a filter is $\uparrow a = \{b \in A : b \geqslant a\}$, but it not always is an ultrafilter.

One way to think of ultrafilters is as logical models or possible worlds (as in the motivation for idea (1) above). Think of the elements of A as propositions. Then the axioms say: (1) if proposition a "implies" proposition b and a is true in model U, then also b is, (2) if propositions a and b are true in model U, also their conjunction $a \wedge b$ is, (3) the tautology 1 is true in the model U, (4) the contradiction 0 is not true in the model U, (5) propositions are bivalent, i.e., either a proposition or its negation are true in the model U. So the collection X of all ultrafilters on A can be seen as the collection of all possible models/worlds.

We now use this tool to construct the embedding. The idea is to map an element $a \in A$ to its 'truth-value profile' across the possible models.

Exercise 2.c (Problem)**.** Let A be a Boolean algebra. Let $X := \{U \subseteq A : U \text{ is an ultrafilter on } A\}$. Define $e : A \to \mathbf{2}^X$ by

$$e(a)(U) := \begin{cases} 0 & \text{if } a \notin U \\ 1 & \text{if } a \in U. \end{cases}$$

$e(a)$ must be a function $X \to 2$; and this definition says how $e(a)$ maps an input $U \in X$ to an output in 2

Show that e is a BA-embedding. *Hint*: You may use the so-called *Boolean Ultrafilter Theorem*: If $a \neq b$ in A, there is an ultrafilter U of A containing one and only one of a and b.

That is one version in which this theorem is usually stated. If you want to read (much) more about this, see, e.g., Davey and Pristley (2002, ch. 10, esp. thm. 10.22).

Now, the last exercise is to finish the proof.

Exercise 2.d (Problem)**.** Prove the desired implication (3)⇒(4) by formalizing the last two steps of the idea (steps 2–3). *Hint*: Think about how valuations can be 'transported along' the embedding e and the projections p_j.

Congratulations: No matter how far you got, you just did some intense mathematics. Be proud!

Exercise 2.e (Philosophical)**.** Possible worlds semantics is one of the success stories of philosophy: A simple theory that still can analyze many important concepts remarkably well—not only meaning but also knowledge, belief, and more. The theory hence found applications also outside of philosophy in linguistics or computer science. According to it, the meaning of a sentence, i.e., the proposition that it expresses

For more background, see Jago (2014, ch. 1).

> may be thought of as a set of possible worlds: the set of worlds in which the sentence expressing the proposition denotes the value true (R. C. Stalnaker 1976, p. 80).

Explore aspects of how the ideas from this chapter relate to possible worlds semantics. Examples of questions that you might investigate are: Do the possible world propositions also form a Boolean algebra (and if so, how)? (If so, lending philosophical plausibility to the proposition motivation of Boolean algebras.) Dually, is the corresponding state-based semantics (built in the preceding exercises) the one where the possible worlds are the (classical) states? What does the logicians' tolerant attitude of working with the whole class of Boolean algebras mean dually for the set of possible worlds? Does this help for the philosophical question of what possible worlds are? Can you think of philosophical criticisms of the view of propositions as sets of possible worlds? How would an alternative look like? What change to the concepts of this chapter would this require? But feel free to also focus on other aspects.

2.6 Notes

Section 2.2 loosely follows Priest (2008, ch. 1). For material on the other sections, see the beginnings of M. J. Dunn and Hardegree (2001). For a short history of classical logic, see, e.g., the introduction of Hodges (1983).

3 Many-valued logic

Now that we've seen classical logic, we will, in the remainder of these notes, look at different logics—hence referred to as *non-classical logics*. The motivation for considering such logics typically arises either in situations where classical logic doesn't seem to be appropriate (e.g., in paradoxes) or when using a different interpretation of what the logical formulas mean (e.g., in intuitionistic logic).

Here we will stay neutral on the status of classical logic: whether it (or any other logic) is the one correct logic and any data to the contrary can be explained away (*logical monism*) or whether there are several correct logics, maybe suitable for different situations (*logical pluralism*). However, as usual in logic and mathematics, we will always use classical reasoning in our metalanguage in which we talk about the various non-classical logics.

For more on this, see, e.g., J. Beall and Restall (2005) or G. Russell (2021).

Instead, we consider the motivation for various logics, study them as a formal framework, and finally discuss some arguments assessing how well that formal logic fares with respect to its motivation. I won't pass judgment on the ultimate correctness of a logic, but this shouldn't stop you from doing so if you find compelling reasons.

It's worth meditating on this statement. For more inspiration, read Sider (2010, sec. 3.3).

In this chapter, we start with a class of non-classical logics called *many-valued logics*. They add finitely many new truth-values to the classical truth-values 0 and 1. Here we restrict ourselves to three- or four-valued logics.

Key concepts

- Motivations for further truth-values (vagueness, liar, future, fiction, denotation, presupposition, topic, databases, etc.)
- Many-valued logic (template)
- Strong and weak Kleene, Łukasiewicz
- Logic of paradox, ST logic, FDE
- Solutions to sorites paradox
- Supervaluationism
- Higher-order vagueness

3.1 Motivation: adding a few more truth-values

We collect many different kinds of motivation for adding more truth-values. We just collect them, but don't discuss whether they ultimately succeed as arguments against for non-classical truth-values. (Instead, this is the topic of exercise 3.a.)

3.1.1 Vagueness

We've already met the sorites paradox which crucially involves vague predicates (like heap, bald, tall, etc.). There, we would say that the sentence

10 grains of sand make a heap. (3.1)

is clearly false (i.e., has truth-value 0), while the sentence

1 billion grains of sand make a heap. (3.2)

is clearly true (i.e., has truth-value 1). However, it's unclear for intermediate claims, like

100 grains of sand make a heap. (3.3)

We neither would say it clearly is true nor would we say that it clearly is false. That is, the truth-value of the sentence is undetermined.

Or choose another number than 100 if you feel differently.

So it is suggestive to add a third truth-value, say i, which stands for undetermined: thus, sentence (3.1) has truth-value 0, sentence (3.2) has truth-value 1, and sentence (3.3) has truth-value i.

However, if we introduce such a new truth-value, we need to say what its logic is: How does the new truth-value interact with old ones, e.g., what is the truth-value of the implication occurring in the sorites paradox: 'If 100 grains of sand don't make a heap, then 101 grains of sand don't'? This we'll do in the next section, but first let's consider some other motivation for adding new truth-values.

3.1.2 Indeterminacy

We gather some more situations where common-sense would say sentences are neither true nor false—because there is no fact to the matter which would determine their (classical) truth-value.

Future contingents: We (like to) think that the future is undetermined:

that a sentence about the future like

Tomorrow, I'll do sports. (3.4)

is neither true nor false since I could still decide either way—there presently simply are no facts that would make the sentence true or false.

For a discussion, see Priest (2008, sec. 7.9).

Fiction: Some might think that sentences from fiction, like "Sherlock Holmes is a detective" are neither true nor false: again, there is no fact about *our* world which would make them true or false; for starters, 'Sherlock Holmes' doesn't refer to a real entity. Maybe even more compellingly, a sentence like

Sherlock Holmes has a mole on his left leg. (3.5)

This example is from Sider (2010, sec. 3.4). Also see Priest (2008, sec. 7.8)

is neither true nor false since not even within the Sherlock Holmes stories written by Sir Arthur Conan Doyle is this determined.

Denotation failure. Even if sentences are about our world, they can still fail to refer to an entity—which is known as denotation failure. A famous example is due to B. Russell (1905a):

The present king of France is bald. (3.6)

The fact that 'bald' is a vague predicate is not relevant here.

There is no present king of France, so that term doesn't denote anything. Hence it doesn't seem right to either say that the sentence is true or false. (Though there are theories, like Russell's, that would take such sentences as false 'by default'.)

For more, see Priest (2008, sec. 7.8).

Failed presuppostions. So, usually, we assume that the terms of a sentence actually denote entities—and this may fail sometimes. But there also are other forms of presuppositions—which also may fail. For example,

Jack stopped smoking. (3.7)

It presupposes that Jack did smoke in the past. So if Jack is a non-smoker, the sentence neither seems true (Jack didn't smoke to start with), nor does it seem false (Jack doesn't smoke right now).

For more, see Sider (2010, sec. 3.4).

Meaninglessness. Maybe some are inclined to call a sentence with a failed denotation or presupposition meaningless—or at least 'meaning-incomplete'. Another kind of 'meaningless' (or semantic nonsense) is

provided by category mistakes: a famous example is Noam Chomsky's

Colorless green ideas sleep furiously. (3.8)

A category mistake *is attributing a property to a thing that, by its very nature (i.e., the category of things it belongs to) cannot have this property.*

Although it is difficult to precisely define 'meaningless' or 'nonsensical', such sentences also neither seem true nor false—it just doesn't make sense to call them true or false.

Off-topic. In a discussion, if we—intuitively speaking—want to wholeheartedly call a sentence 'true', we not only want that it says something correct about our world but also that it is about the topic of the discussion. For example, in a discussion about our solar system, we only wholeheartedly want to call

For an early logical analysis, see Bochvar (1938, 1981).

The earth revolves around the sun. (3.9)

true but not

The smallest prime number is 2. (3.10)

If someone responds to (3.9) with (3.10), we'd respond by: yes, that's correct, but has nothing to do with what we're talking about. (Although, at least in philosophy, one sharply distinguishes truth and topic, so many philosophers would object to this intuition.) So if our 'beefed up' true means *true and on topic* and our 'beefed up' false means *false and on topic*, we have (at least) a third option, namely off-topic. This again motivates a third 'undetermined' truth-value.

See J. c. Beall (2016). For an topic or 'aboutness' interpretation with four (or more) truth-values, see J. M. Dunn (1966, 2019).

All of these sources of indeterminacy need—and have received—a careful philosophical analysis. Some starters are found in the quoted literature. For us now, it's only important that they deliver some intuitive motivation for adding a third truth-value i for "undetermined" or "neither true nor false".

3.1.3 Liar paradox

We've already seen the liar paradox: The liar sentence "This sentence is false" is true iff it is false. This allowed (at least) two conclusions that avoid contradiction: either it is *both true and false* or it is *neither true nor false*. The second seems less outrageous—especially after now having seen much motivation for a third *neither true nor false* truth-value. However, as already seen, it faces the revenge paradox ("This sentence is either false or neither true nor false"). This motivates considering as a third truth-value i

representing *both true and false*, and taking the liar sentence to have this truth-value.

3.1.4 Databases: incomplete and inconsistent data

Consider a computer—or artificial intelligence (AI)—which has access to a database, say about goods stored in a warehouses. It gets user queries asking whether the query is true or false: e.g., are there five packs of pasta? The database is not very well-maintained: It could be incomplete (e.g., no information about pasta) or inconsistent (someone entered that there are 2 packs of pasta while someone else entered 5 packs). Belnap (2019a,b) asks: How should the computer (reason to) respond to these queries?

In the incomplete case, it would be natural to say the query about packs of pasta is *neither true nor false* (according to the database). In the inconsistent case, it is tempting to reply *both true and false*: it is true according to the database that there are 5 packs (someone entered 5 packs), but it also is false (someone entered 2 packs).

Then it is an important matter to find the right logic to reason with these truth-values for more complex queries. (Or drawing conclusions about the current query from other, not directly related queries.) In particular, it cannot just be classical logic: since, then, from $p \wedge \neg p$ any sentence q follows. So once the database would be 'locally inconsistent' (e.g., inconsistent about the packs of pasta but otherwise fine), it would be made 'globally inconsistent' by classical logic since it answers 'true' to any query.

3.2 Formal logics: many-valued logics

Many-valued logics generalize classical logic quite straightforwardly, based on a truth-value semantics. (Recall from chapter 2 that a truth-value semantics is a special case of an algebraic semantics using a single algebraic structure of truth-values.)

See e.g. Priest (2008, p. 7.2).

Template 3.1 (Many-valued logic)**.**

- Choose a finite set of truth-values $\mathsf{T} \supseteq \{0, 1\}$ extending the classical ones 0 and 1.
- Define a notion of valuation as a function $v : \mathsf{P} \to \mathsf{T}$ that extends to all formulas via truth-tables (that define the truth-functions on T interpreting the connectives).
- To define validity, first define a set of *designated values* $\mathsf{D} \subseteq \mathsf{T}$. (In classical logic, $\mathsf{D} = \{1\}$.)

- Then consequence $\Gamma \models \varphi$ is defined as preservation of designated values: for any valuation v such that $v(\psi) \in D$ for all $\psi \in \Gamma$, also $v(\varphi) \in D$.
- A generalization is to consider two sets of designated values $D_p, D_c \subseteq T$ ('p' for premises and 'c' for conclusion), and define $\Gamma \models \varphi$ as: for any valuation v, if $v(\psi) \in D_p$ for all $\psi \in \Gamma$, then $v(\varphi) \in D_c$. If $D_p \neq D_c$ one speaks of a *mixed consequence* relation.

If L is a logic obtained this way, we say it is an n-*valued logic* if T has n elements. And we call its valuations (with the intended way to extend them all formulas) L-*valuations*.

We now fill in this template with concrete examples.

3.2.1 Strong Kleene, weak Kleene, and Łukasiewicz

The *strong Kleene logic* K_3^s is obtained from the template by taking $T = \{0, 1, i\}$, $D = \{1\}$, and the truth-tables

$\neg$	
1	0
i	i
0	1

$\wedge$	1	i	0
1	1	i	0
i	i	i	0
0	0	0	0

$\vee$	1	i	0
1	1	1	1
i	1	i	i
0	1	i	0

$\rightarrow$	1	i	0
1	1	i	0
i	1	i	i
0	1	1	1

and $\top$ (resp., $\bot$) is always evaluated to 1 (resp., 0), and $\varphi \leftrightarrow \psi$ is evaluated as $(\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$. The *weak Kleene logic* K_3^w is the same except that any truth-function maps to i as soon as one argument is i. The *Łukasiewicz logic* $Ł_3$ is the same as K_3^s except that $i \rightarrow i = 1$ (instead of i).

A notational subtlety: A function $v : P \rightarrow \{0, 1, i\}$ (i.e., a three-valued valuation of propositional atoms) could be either a K_3^s-valuation or a K_3^w-valuation or a $Ł_3$-valuation, depending on which truth-tables we use to extend it to all formulas. Usually, which to use is clear from context, but if we want to stress it, we may write $v_{\mathsf{K}_3^w}$, $v_{Ł_3}$, etc.

The intuition behind these truth-tables is the following.

Weak Kleene. Here i is thought of as either non-sense/meaningless or as off-topic (cf. section 3.1.2). As soon as one part of the sentence is non-sense or off-topic, the whole sentence is non-sense or off-topic, and on the remaining classical values it acts like classical logic.

Strong Kleene and Łukasiewicz. Here i is thought of as neither true nor false. The idea for, e.g., $0 \wedge i = 0$ is this: currently, i is undetermined, but no matter how i eventually gets determined—i.e., no matter which classical truth-value it eventually gets—the conjunction $0 \wedge i$ always is 0

(since $0 \wedge 0 = 0$ and $0 \wedge 1 = 0$). On the other hand, $i \wedge 1 = i$ because: if i gets determined as 0, then $0 \wedge 1 = 0$, but if i gets determined as 1, then $1 \wedge 1 = 1$; so a determination of i doesn't fix the classical truth-value of $i \wedge 1$.

Exercise 3.2. Go through more values of the truth-table and check this intuition.

So the general intuition is this: Consider an n-ary connective $\star$ like $\neg$, $\wedge$, $\vee$, or $\rightarrow$. Then $\star(a_1, \ldots, a_n) = 1$ (resp., $= 0$) if, for every replacement of i among the $a_1, \ldots, a_n$ by classical truth-values, the result, after applying the classical truth-table for $\star$, is 1 (resp. 0). Otherwise $\star(a_1, \ldots, a_n) = i$.

The only difference between K_3^s and Ł$_3$ is how to disambiguate this intuitive rule in the case $i \rightarrow i$. If we determine i as 1, then $1 \rightarrow 1 = 1$, so $i \rightarrow i$ certainly can become true. Now, Łukasiewicz says, if we determine i as 0, also $0 \rightarrow 0 = 1$, so on either determination $i \rightarrow i$ becomes 1, so $i \rightarrow i = 1$. But strong Kleene says: we could determine the two occurrences of i *independently*, and if the first is determined to 1 and the second determined to 0, then $1 \rightarrow 0 = 0$ also can become false, so $i \rightarrow i = i$.

This may seem annoyingly pedantic, but wait for it (exercise 3.4 below)!

Two more comments on valuations: First, the truth-tables for the Boolean connectives can also be given another interpretation in terms of min and max as in classical logic: see section 3.2.4 below. Second, in classical logic, we said that a valuation is like a possible world in the sense of a complete and consistent collection of atomic facts—which we called a classical state. A three-valued valuation (with i interpreted as undetermined) can be seen as a generalization that drops the completeness assumption: they now are possibly incomplete but still consistent collections of atomic facts—sometimes called *situations*. If we also drop the remaining consistency assumption, we get to possibly incomplete and possibly inconsistent collections of atomic facts—like databases. These will be the valuations in the four-valued logic FDE below.

For the rest of this subsection, let's do some calculations to get a feel for these logics. First, some differences and commonalities with classical logic.

Example 3.3.

1. Unlike classical logic, the law of excluded middle, $p \vee \neg p$, is not valid in any of the three logics $K_3^s, K_3^w, Ł_3$: give p the value i (make sure you know how to do this formally), then $\neg p$ also has value i, so $p \vee \neg p$ has value $i \vee i = i$, which is not designated.

2. In K_3^s we still have contraposition like in classical logic: $\neg q \rightarrow \neg p \vDash_{K_3^s} p \rightarrow q$. Let v be a strong Kleene valuation with $v(\neg q \rightarrow \neg p) = 1$.

Then, looking at the table for $\to$, either $\nu(\neg p) = 1$ or $\nu(\neg q) = 0$. If $\nu(\neg p) = 1$, then $\nu(p) = 0$, so $\nu(p \to q) = 1$ also is designated. If $\nu(\neg q) = 0$, then $\nu(q) = 1$, so $\nu(p \to q) = 1$ also is designated.

⌟

The difference between strong Kleene and Łukasiewicz in their treatment of $i \to i$ may seem small, but the effects are huge:

Exercise 3.4. 1. Show that $\models_{Ł_3} p \to p$ while $\not\models_{K_3^s} p \to p$.

2. Let $\nu : P \to \{0, 1, i\}$ be the valuation mapping each p to i. In K_3^s, if φ is a formula without $\top$ and $\bot$, then $\nu(\varphi) = i$. In $Ł_3$, if φ is a Boolean formula without $\top$ and $\bot$, then $\nu(\varphi) = i$.

3. Conclude that, in fact, K_3^s doesn't have any ($\top$ and $\bot$ free) validities, i.e., for any formula φ without $\top$ and $\bot$, we have $\not\models_{K_3^s} \varphi$.

The difference between K_3^s and $Ł_3$ also shows up in their treatment of the conditional:

Exercise 3.5. 1. Show that in K_3^s, $\neg\varphi \vee \psi$ is equivalent to $\varphi \to \psi$ in the sense of having the same value under any strong Kleene valuation. (In particular, they mutually entail each other.) Thus, in strong Kleene logic, the conditional $\to$ is the material conditional and we also can, without loss of generality, work in the Boolean language $\mathcal{L}_{\text{bool}}$.

2. Conclude using exercise 3.4 (3) that in $Ł_3$, $p \to q$ cannot be defined using Boolean connectives: i.e., for any Boolean sentence φ without $\top$ and $\bot$, there is a Łukasiewicz valuation ν such that $\nu(\varphi) \neq \nu(p \to q)$ (in fact, $p \to q \not\models_{Ł_3} \varphi$).

From Sider (2010, ex. 3.6).

3.2.2 Logic of paradox LP

Now we see that the choice of designated values D really matters: The *logic of paradox* LP is defined again using $T = \{0, 1, i\}$ with the same truth-tables as strong Kleene logic, but now $D = \{1, i\}$ instead of $\{1\}$. In particular, an LP-valuation is the same as a K_3^s-valuation, so we keep using the latter term. Despite this subtle change in D, this yields quite a different logic.

Due to Priest (1979).

The interpretation of i now is as *both true and false*—as it occurs, e.g., as one response to the liar paradox. Thus, 1 is thought of as *true and not false* and 0 as *false and not true*. Coincidentally, this interpretation of i precisely works with the strong Kleene truth-tables (hence the definition of LP): We

only need to consider $\neg, \wedge, \vee$ since $\rightarrow$ is definable (exercise 3.5 (1)). For example, we have $\neg i = i$ because if a sentence p is both true and false, then $\neg p$ is true (because p is false) and false (because p is true), so $\neg p$ is i again. And we have $1 \wedge i = i$ because if p is true and not false and q is both true and false, then $p \wedge q$ is true (since both p and q are true) and false (since q is false), so $p \wedge q$ is i. Similarly for other combinations (check some more as an exercise).

This interpretation also motivates the choice of D: Commonly, we think of validity as truth-preservation. But then, in this setting, both 1 and i need to be preserved, because both contain truth!

To be a 'logic of paradox', i.e., a logic that can deal with contradictions, these contradictions should not trivialize the logic as they do in classical logic. So it shouldn't satisfy the *ex falso quodlibet* rule: $p, \neg p \vDash q$ which is valid in classical logic. Indeed, consider a valuation v giving p value i and q value 0. So p is a contradiction, i.e., both true and false, and q is false and not true. Then $v(p) = i = v(\neg p)$ is designated (i.e., in D), but $v(q)$ is not. So $p, \neg p \nvDash_{\mathsf{LP}} q$, as needed.

That's Latin for 'from falsehood, anything [follows]'.

Note that this is only due to the small change to D: in strong Kleene we still have $p, \neg p \vDash_{\mathsf{K}_3^s} q$: trivially so, because there is no strong Kleene valuation v such that $v(p), v(\neg p) \in D$. For the same reason, we now, unlike in K_3^s, have the law of excluded middle $\vDash_{\mathsf{LP}} \varphi \vee \neg\varphi$ (no valuation makes this undesignated, i.e., 0).

Common terminology is to say: a logic L is *paraconsistent* if $p \wedge \neg p \nvDash_{\mathsf{L}} q$ (i.e., $p \wedge \neg p$ doesn't imply everything). And L is *paracomplete* if $q \nvDash_{\mathsf{L}} p \vee \neg p$ (i.e., $p \vee \neg p$ doesn't follow from everything). So LP is complete (i.e., not paracomplete) but paraconsistent, while K_3^s is consistent (i.e., not paraconsistent) but paracomplete.

However, this also comes at a cost: modus ponens ($p, p \rightarrow q \vDash q$), a logical law we hold rather dearly, fails:

Because of this, one might change the truth-table for $\rightarrow$ slightly and obtain a version of LP called RM_3: see Priest (2008, par. 7.4.6–8).

Exercise 3.6. Show $p, p \rightarrow q \nvDash_{\mathsf{LP}} q$, while $p, p \rightarrow q \vDash_{\mathsf{K}_3^s} q$

Nonetheless, one sense in which LP is rather well-behaved is that it has exactly the same validities as classical logic: We work toward that result in the next exercise.

This is an example that logics can have the same validities (i.e., $\emptyset \vDash \varphi$ is the same) but different consequence relations (i.e., $\Gamma \vDash \varphi$ may differ)!

Exercise 3.7 (This also is an end-of-chapter exercise)**.** 1. First, a helpful concept: If $v, w : \mathsf{P} \rightarrow \{0, 1, i\}$ are two K_3^s-valuations (so also LP-

valuations), we define $v \leqslant w$ (w *refines* v) if, for all $p \in \mathsf{P}$,

$$\begin{array}{ll} \text{If } v(p) = 0, \text{ then } w(p) = 0 & \text{, and} \\ \text{if } v(p) = 1, \text{ then } w(p) = 1. & \end{array}$$

So w has at least as much classical information as v: it may make more propositional atoms decided, but it agrees with v on the atoms that v decided on.

Show that this extends to all sentences φ (in K_3^s or LP): if $v(\varphi) = 0$ (resp., $= 1$), then $w(\varphi) = 0$ (resp., $= 1$). As a slogan: In K_3^s and LP, classical truth-values are stable under valuation refinement. (Note: this is not true for $Ł_3$-valuations!)

2. Another helpful concept: A K_3^s-valuation (so also an LP-valuation) $v : \mathsf{P} \to \{0, 1, i\}$ is *classical* iff $v(p) \in \{0, 1\}$ for all $p \in \mathsf{P}$. So it may also be considered as a classical valuation $v : \mathsf{P} \to \{0, 1\}$. Show: If v is a classical K_3^s-valuation and φ a sentence, then $v(\varphi)$ determined according to the K_3^s-truth-tables is identical to $v(\varphi)$ determined according to the truth-tables of classical logic—in short: $v_{\mathsf{K}_3^s}(\varphi) = v_{\mathsf{CL}}(\varphi)$.
3. Conclude that, for all propositional formulas φ, we have $\models_{\mathsf{LP}} \varphi$ iff $\models_{\mathsf{CL}} \varphi$.

3.2.3 ST Logic

Now we see an example of a mixed consequence relation, i.e., choosing two separate sets of designated values for premises and conclusion, respectively.

See Cobreros et al. (2012, 2015, 2020).

The logic ST is again defined using $T = \{0, 1, i\}$ with the same truth-tables as strong Kleene logic, but now with $D_p = \{1\}$ (the designated values of K_3^s) and $D_c = \{1, i\}$ (the designated values of LP). So, like LP-valuations, also ST-valuations simply are K_3^s-valuations.

The interpretation of i is again as both true and false. Then, since 1 is 'true and not false', being D_p-designated can be thought of as being *strictly* true. Being D_c-designated means containing truth, so it can be thought of as being *tolerantly* true. Thus, $\Gamma \models_{\mathsf{ST}} \varphi$ means that the strict truth of the premises Γ implies the tolerant truth of the conclusion φ—hence ST-logic.

What does the consequence relation $\models_{\mathsf{ST}}$ on this 'fusion' of K_3^s and LP look like? It precisely is classical logic!

Exercise 3.8 (This also is an end-of-chapter exercise)**.** Use the helpful concepts about K_3^s-valuations from exercise 3.7 to conclude that: For any set of propositional formulas Γ and any propositional formula φ, we have

$$\Gamma \models_{\mathsf{ST}} \varphi \text{ iff } \Gamma \models_{\mathsf{CL}} \varphi.$$

So we could add the ST-semantics as one of the equivalent semantics for classical logic from theorem 2.19.

But then what's the point of ST-logic? The point is that it provides possible solutions to the sorites and the liar paradox once the vocabulary is added to state the paradoxes. We get to this in section 3.3.

3.2.4 FDE

The last many-valued logic that we'll consider is *first-degree entailment* (FDE). Belnap (2019a,b) motivated and developed this logic in the 1970s as "how a computer should think": how a computer should reason (to answer queries) given the data from a possibly inconsistent and incomplete database (as described in section 3.1.4). Here is the formal definition.

See Priest (2008, ch. 8). The name comes from relevance logics (which we'll see later). These are various logics for the conditional $\rightarrow$ (or 'entailment') and FDE is obtained if no nesting of the conditional is allowed (i.e., only first-degree): $\varphi \vDash \psi$ iff $\varphi \rightarrow \psi$ is valid in the appropriate relevance logic.

The logic FDE is four-valued: $\mathsf{T} = \{0, 1, \mathfrak{b}, \mathfrak{n}\}$ where $\mathfrak{b}$ is interpreted as *both true and false* and $\mathfrak{n}$ as *neither true nor false* (and 1 is true and not false while 0 is false and not true). The designated values are $\mathsf{D} = \{1, \mathfrak{b}\}$, so validity is preservation of truth. The truth-tables are (to be explained below)

$\neg$	
1	0
$\mathfrak{b}$	$\mathfrak{b}$
$\mathfrak{n}$	$\mathfrak{n}$
0	1

$\wedge$	1	$\mathfrak{b}$	$\mathfrak{n}$	0
1	1	$\mathfrak{b}$	$\mathfrak{n}$	0
$\mathfrak{b}$	$\mathfrak{b}$	$\mathfrak{b}$	0	0
$\mathfrak{n}$	$\mathfrak{n}$	0	$\mathfrak{n}$	0
0	0	0	0	0

$\vee$	1	$\mathfrak{b}$	$\mathfrak{n}$	0
1	1	1	1	1
$\mathfrak{b}$	1	$\mathfrak{b}$	1	$\mathfrak{b}$
$\mathfrak{n}$	1	1	$\mathfrak{n}$	$\mathfrak{n}$
0	1	$\mathfrak{b}$	$\mathfrak{n}$	0

and $\bot$ and $\top$ always evaluate to 0 and 1, respectively, and the conditional $\varphi \rightarrow \psi$ is the material conditional $\neg\varphi \vee \psi$ and the biconditional $\leftrightarrow$ is defined as usual with $\rightarrow$ and $\wedge$.

A motivation for the choice of truth-values is given by databases (section 3.1.4): A database $\mathfrak{B}$ determines a valuation ν as follows

$\nu(p) = 1$	p is true according to $\mathfrak{B}$ and p isn't false according to $\mathfrak{B}$
$\nu(p) = \mathfrak{b}$	p is true according to $\mathfrak{B}$ and p is false according to $\mathfrak{B}$
$\nu(p) = \mathfrak{n}$	p isn't true according to $\mathfrak{B}$ and p isn't false according to $\mathfrak{B}$
$\nu(p) = 0$	p isn't true according to $\mathfrak{B}$ and p is false according to $\mathfrak{B}$.

Further, an intuitive interpretation of the truth-table is as follows. (This also provides another motivation for the Boolean fragment of the preceding logics.)

We can order the truth-values $\mathsf{T} = \{0, \mathfrak{b}, \mathfrak{n}, 1\}$ by 'how true they are' as in figure 3.1. The idea is:

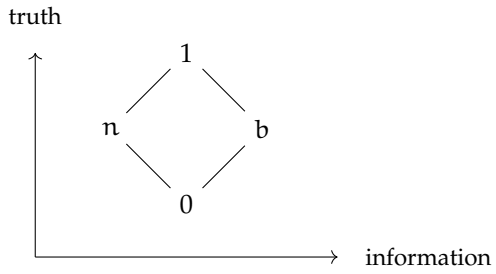

Figure 3.1: Ordering of the four truth-values.

- 0 is least true because it is 'completely false', and 1 is greatest since it is 'completely true'.
- n is more true than 0 because it is not false, but it still is less true than 1 since it also is not true.
- b is more true than 0 because it is true, but it still is less true than 1 since it also is false.
- b and n are not comparable: over-determined is not more true than under-determined and vice versa.

Now the truth-functions are defined as follows: For negation, we expect, as in LP, that $\neg\varphi$ contains truth iff φ contains falsity. So, e.g., $\neg\varphi$ is b iff $\neg\varphi$ contains both truth and falsity iff φ contains both falsity and truth iff φ is b. For conjunction, generalizing classical logic, $a \wedge b$ simply is the minimum $\min(a, b)$ (or, more correctly, the greatest lower bound) in the order of T from figure 3.1. For disjunction, $a \vee b$ is the maximum $\max(a, b)$ (or, more correctly, the least upper bound) in that order. Exercise 3.d gives more explanation why exactly to choose these operations.

As promised, this also works for the Boolean connectives of K_3^s (and hence also for Ł$_3$, LP, and ST). There we only have one more value i in addition to the classical 0 and 1 and it could be, depending on the intended interpretation, b or n. After discarding one, the order of $\mathsf{T} = \{0, b, n, 1\}$ from figure 3.1 becomes that of figure 3.2. Then $\wedge$ and $\vee$ of the strong Kleene truth-tables again simply are min and max in that order.

Maybe you thought—like Belnap (2019a)—that there also is *another* natural ordering on the four truth-values: namely, by information. The value n carries the least amount of information (in fact, no or 'incomplete' information), the values 0 and 1 contain more, but incomparably much information (they each provide complete and consistent information), and

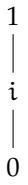

Figure 3.2: Ordering of the three truth-values (obtained from 3.2 by either taking $\mathfrak{i} = \mathfrak{b}$ or by taking $\mathfrak{i} = \mathfrak{n}$).

the value $\mathfrak{b}$ contains a maximal amount of information (in fact, too much to be consistent). This order is given as the Hasse diagram in figure 3.1 but now read from left-to-right. This order is not needed to compute truth-values, but yields a rather pleasing interpretation of the 'truth-value diamond'. And it simplifies certain notions: e.g., the notion of refinement from exercise 3.7. Given two valuations $v, w : \mathsf{P} \to \{0, 1, \mathfrak{i}\}$, we have that w refines v (i.e., $v \leqslant w$) iff for all $p \in \mathsf{P}$, $v(p) \leqslant_I w(p)$ where $\leqslant_I$ is the information ordering.

We see this again in chapter 9.

To practice computations in this logic, show that the De Morgan laws remain valid.

Exercise 3.9. Show $\neg(\varphi \wedge \psi) \vDash_{\mathsf{FDE}} \neg\varphi \vee \neg\psi$ and $\neg\varphi \vee \neg\psi \vDash_{\mathsf{FDE}} \neg(\varphi \wedge \psi)$.

3.3 Assessment

Having seen the formal definitions of several many-valued logics, it now is high time to see how they deliver on their motivations. We'll focus here on vagueness.

3.3.1 Sorites paradox

To state the sorites paradox, we need to fix some vocabulary. To recall the paradox, write

$$p_n = \text{ } n \text{ grains of sand don't make a heap.}$$

The paradox then is (for some large N): as premises, both 'p_1' and 'if p_j, then p_{j+1}' (for $j = 1, \ldots, N-1$) seem plausible, and they seem to imply (by modus ponens) that p_N, but p_N is false.

Let's see how this is modeled formally by the many-valued logics. We take valuations (of the different logics) as potential models for the sorites situation. Obviously not any valuation will be a good model: the classical

valuation setting every p to 0 is not paradoxical whatsoever, but it also doesn't capture the idea that 'p_1' and 'if p_j, then p_{j+1}' seem plausible (since it says that p_1 is false).

So, in classical logic, we could straightforwardly capture this constraint by saying that only those valuations are 'good' that satisfy $\nu(p_1) = 1$ and $\nu(p_j \to p_{j+1}) = 1$. But then, since classical logic satisfies modus ponens ($p, p \to q \vDash_{\mathsf{CL}} q$), any such 'good' valuation will satisfy $\nu(p_N) = 1$, so it cannot capture the intuition that p_N should be false—this precisely is the paradox (now stated formally).

The strong Kleene solution to vagueness.

One response is to consider three-valued valuations $\nu : \mathsf{P} \to \{0, 1, i\}$ thinking of i as neither true nor false. The 'good' valuations should still satisfy $\nu(p_1) = 1$, but we now only require $\nu(p_j \to p_{j+1}) \in \{1, i\}$, whence 'seems plausible' is understood as 'is not false' (a weak form of plausibility). The motivation might be that eventually in the sequence $p_1, \ldots, p_N$ the p_j's are neither true nor false (and, yet further, false), so eventually we should have a conditional $p_j \to p_{j+1}$ going from 1 to i, and $1 \to i = i$ (in both the strong Kleene, weak Kleene, and Łukasiewicz truth-tables). But now, e.g., a valuation ν with $\nu(p_1) = 1$, $\nu(p_2) = \ldots = \nu(p_{N-1}) = i$, and $\nu(p_N) = 0$ is a 'good' model of the situation in this above weak sense (since $\nu(p_1) = 1$ and $\nu(p_j \to p_{j+1}) \in \{1, i\}$) and still satisfies $\nu(p_N) = 0$, hence avoiding paradox.

The ST *solution to vagueness.*

However, one might object: 'not false' is too weak a sense to capture 'plausible'. Here is another suggestion following ST logic (Cobreros et al. 2015). Represent the natural language rule 'if p_n, then p_m' not as a conditional $p_n \to p_m$ but rather as the statement

$$q_{n,m} = n \text{ and } m \text{ grains of sands are indiscriminable with respect to making a heap.}$$

and capture its content instead by requiring that a *good* valuation satisfies

$$\nu(q_{n,m}) = \begin{cases} 0 & \text{if } \nu(p_n) = 0 \text{ and } \nu(p_m) = 1 \text{ or vice versa} \\ 1 & \text{otherwise.} \end{cases} \tag{3.11}$$

That is, a three-valued valuation ν is a 'good' model of the sorites situation if it makes true $q_{n,m}$ whenever p_n and p_m are close in truth-values, and false otherwise (so never leaves $q_{n,m}$ undetermined). To analyze the paradox, we want to understand the ST logic when restricting to good valuations:

Define $\Gamma \models^*_{\mathsf{ST}} \varphi$ by: for all good valuations v (i.e., ST-valuations satisfying equation 3.11), if $v(\psi) = 1$ for all $\psi \in \Gamma$, then $v(\varphi) \neq 0$.

Then the individual steps of the sorites argumentation (i.e., the individual modus ponens applications) are valid—just as we intuitively think:

$$p_j, q_{j,j+1} \models^*_{\mathsf{ST}} p_{j+1} \qquad \text{(for } j = 1, \ldots, N-1\text{)}.$$

(Proof: if v is a good valuation with $v(p_j) = 1$ and $v(q_{j,j+1}) = 1$, then $v(p_{j+1})$ cannot be 0 since otherwise $v(q_{j,j+1}) = 0$.) However, paradox is avoided because $\models^*_{\mathsf{ST}}$ is not transitive:

$$p_1, q_{1,2}, q_{2,3} \not\models^*_{\mathsf{ST}} p_3$$

(Proof: take $v(p_1) = 1$, $v(p_2) = i$, $v(p_3) = 0$ and $v(q_{1,2}) = 1 = v(q_{2,3})$, which is a good valuation making true the premises but making false the conclusion.)

Exercise 3.10. Compare these two suggested solutions: Which do you think is more convincing? Which advantages and disadvantages do they have?

3.3.2 Supervaluationism

See Sider (2010, sec. 3.4.5).

It's natural to regard a sentence like 'This is a heap' to be (definitely) true, if on all ways the vague word 'heap' could reasonably be made precise, the sentences is true. For example, if we're pointing at a collection of 1 billion grains of sand, this is the case because on all reasonable precisifications of 'heap'—e.g., 'more than 1000 grains', 'more than 10.000 grains', etc.—the sentence is true. So truth of vague sentences means truth in all precisifications of the sentence.

We've already seen this idea in the motivation for the strong Kleene truth-tables: the truth-function is 1 (resp., 0) if, on any way of making the i's in the arguments determined, classical reasoning yields 1 (resp., 0), and otherwise it is i. Supervaluationism takes this idea further: not only to determine truth-tables, but—as motivated in the preceding paragraph—also to apply to all formulas. Formally:

- A classical valuation $w : \mathsf{P} \to \{0, 1\}$ is a *precisification* of (or a *way of determining*) a three-valued valuation $v : \mathsf{P} \to \{0, 1, i\}$ if w refines v. So whenever v assigns a classical truth-value to p, w assigns the same classical value, but w also assigns classical values to those p where v is undetermined.

- Given a three-valued valuation $\nu : \mathsf{P} \to \{0, 1, i\}$, a formula φ is *supertrue* (resp., *superfalse*) with respect to ν iff $w_{\mathsf{CL}}(\varphi) = 1$ (resp., $w_{\mathsf{CL}}(\varphi) = 0$) for any precisification w of ν. If so, we write $\nu \models^{!1} \varphi$ (resp., $\nu \models^{!0} \varphi$).
- Supervaluational consequence $\Gamma \models_S \varphi$ is defined as: for any three-valued valuation ν, if $\nu \models^{!1} \psi$ for all $\psi \in \Gamma$, then $\nu \models^{!1} \varphi$.
- This 'global' definition of $\models_S$ can equivalently be given 'locally' (you could do that as an exercise): For all three-valued valuations ν, for all precisification w of ν, if $w(\psi) = 1$ for every $\psi \in \Gamma$, then $w(\varphi) = 1$.

This does indeed result in a logic different than Kleene's:

Example 3.11. We've seen that $\not\models_{\mathsf{K}_3^s} p \vee \neg p$ (on any strong Kleene valuation ν with $\nu(p) = i$). However, the law of excluded middle is supervaluationally valid, i.e., $\models_S p \vee \neg p$. To see this, let ν be a three-valued valuation and show $\nu \models^{!1} p \vee \neg p$. So let w be a precisification of ν and show $w_{\mathsf{CL}}(p \vee \neg p) = 1$. But this is the case, since $p \vee \neg p$ is a classical tautology, i.e., true under any classical valuation (and w is a classical valuation). ⌟

In fact, this shows that any classical tautology φ is supervaluationally valid (just replace $p \vee \neg p$ by φ). Stronger yet, supervaluational consequence is just classical consequence:

Exercise 3.12. Show: $\Gamma \models_S \varphi$ iff $\Gamma \models_{\mathsf{CL}} \varphi$.

The upshot is: Supervaluationism is a well-motivated logic and semantics to deal with vague sentences. The fact that its consequence relation—i.e., its notion of valid arguments—coincides with classical logic may be interpreted as showing that vagueness doesn't force us to give up classical logic.

Also, supervaluationism may be seen as a criticism of the strong Kleene solution of vagueness: Only supervaluationism but not strong Kleene fully captures the above compelling principle that correct reasoning with vague sentences means 'correct reasoning under any precisification'.

Remark 3.13. However, classical logic and the supervaluationist logic come apart if we consider so-called multi-conclusion consequence: $\varphi \vee \psi \models \varphi, \psi$ holds in classical logic because any valuation making true $\varphi \vee \psi$ will make true some element of $\{\varphi, \psi\}$. But $p \vee \neg p \not\models_S p, \neg p$ since a valuation ν with $\nu(p) = i$ will make supertrue $p \vee \neg p$ but it will neither make supertrue p nor $\neg p$. (They also come apart once 'supertrue' can be expressed in the object language, see the next subsection.)

So why don't we call it a day and take supervaluationism as the correct logic of vagueness? Well, there still are problems: with so-called 'higher-order vagueness'.

3.3.3 Borderline cases and higher-order vagueness

Another criticism of three-valued approaches to vagueness is given by 'border-border-line' cases: The motivation for the truth-value $\mathfrak{i}$ was that, in a long sorites sequence $p_1, \ldots, p_N$, there is no clear-cut boundary between the true sentences and the false sentences: there are borderline sentences in between, with value $\mathfrak{i}$. This, however, produces clear-cut boundaries between 1 and $\mathfrak{i}$ and between $\mathfrak{i}$ and 0—i.e., the borders of the borderline cases. But they seem just as arbitrary as a clear-cut boundary between 0 and 1 (Priest 2008, par. 11.3.8).

Why should it be that 99 grains of sand don't make a heap, but for 100 grains it is undetermined whether they form a heap?

Here are two possible replies. First, one might discard the idea of discrete truth-values all together and rather choose continuously from any number between 0 and 1. This is *fuzzy logic* and we'll consider it in the next chapter. This also has its own problem. So let's first try a more cautious account of our reasoning with border-border-line cases. After all, we don't seem to have much trouble with this in everyday life.

As a general rule, if we want to model our reasoning, a good place to start is how we express it in natural language. In the case of borderline cases, we use the word 'definitely' (or 'clearly', etc.): To say that a sentence p with a vague predicate—like Harry is bald—has truth-value 1, we would say, in ordinary language, 'Harry is *definitely* bald'. To say it has truth-value 0, we say 'Harry is *definitely* not bald'. And to say it has truth-value $\mathfrak{i}$, we would say 'Harry is neither definitely bald nor definitely not bald'. To write this more succinctly, let's introduce a new operator (or 1-ary connective) Δ to our language: where $\Delta\varphi$ should intuitively mean 'definitely φ'. Then we can state these three possibilities as Δp, $\Delta\neg p$, or $\neg\Delta p \wedge \neg\Delta\neg p$, respectively. (The latter we may abbreviate as the indefinite operator ∇p.)

A typical example from the literature.

In the literature, you also find D *instead of* Δ.

So far, this is nothing new. But, while the truth-values exhausted their distinguishability here, we can use the Δ-notation to consider border-borderline cases: If Harry is borderline between 1 and $\mathfrak{i}$, he sits at the border of the definitely bald people and the not definitely bald people, but doesn't 'fully' belong to either group. He neither is definitely 'definitely bald' nor definitely 'not definitely bald': $\neg\Delta\Delta p \wedge \neg\Delta\neg\Delta p$. Harry is an indefinite case of definite baldness.

That the predicate 'definitely bald' is vague, too, is known as second-

order vagueness. And, using the Δ operator, it can be iterated further, so one speaks of higher-order vagueness (Williamson 1999). As a slogan: higher-order vagueness is vagueness in whether there is vagueness (Sider 2010, sec. 3.4.3).

So far we didn't provide a formal logic for the extended language with Δ. And that is a difficult problem. Here are two approaches.

Compare this with the problem of finding a logic/semantics for the operator $\Box$ intuitively standing for 'necessarily'. The solution in this case is modal logic.

First, one might first try to give a semantics to Δ in the context of three-valued logic using the truth-tables

Δ	
1	1
i	0
0	0

But this trivializes higher-order vagueness: $\neg\Delta\Delta p \wedge \neg\Delta\neg\Delta p$ will always have value 0.

Second, one might take Δ to mean supertrue: after all, supertrue means truth on all precisifications, which is a natural way to understand 'definitely'. Typically, one then generalizes the above valuation-semantics to a modal one: the possible states are precisifications and $\Delta\varphi$ is true at a state iff it is true at every state (i.e., supertrue). We don't go into details here, but it is intuitive enough that in this extended language classical (meta-) inferences fail—e.g., the deduction theorem: $p \models_S \Delta p$ (if p is supertrue, then Δp is true everywhere, i.e., supertrue) but $\not\models_S p \rightarrow \Delta p$ (as soon as there is one precisification where p is true and one where it is false, the conditional is false at the former, since the antecedent is true and the consequent false). However, to allow for the possibility $\neg\Delta\Delta p \wedge \neg\Delta\neg\Delta p$ (and further higher-order vagueness) a more careful semantics for Δ is needed (not just as 'global truth'). For further reading, see e.g. Incurvati and Schlöder (2021).

If we want to conclude this chapter with a moral, it might be this: We don't have a problem when reasoning with vague predicates in everyday life (we do this all the time). But we encounter all kinds of problems when we actually want to model that reasoning in a general and formal logic. (And having such a logic arguably is not just philosophically relevant but also practically when, e.g., describing to a computer how to reason with vague predicates.)

3.4 Exercises

Exercise 3.a (Philosophical)**.** Pick a motivation for additional, non-classical truth-values from section 3.1 (vagueness, future contingents, fiction, denotation failure, failed presuppositions, meaninglessness, off-topic, liar, databases). Philosophically discuss whether it constitutes a good argument for adding a third (or also a fourth) truth-value—or whether the motivation can be explained away using only classical truth-values.

You can find inspiration here: (a) Sider (2010, pp. 94–95) for an argument against a truth-value neither true nor false, (b) Priest (2008, sec. 7.9) for Aristotle's argument that future contingents cannot be determined, or (c) Priest (2008, sec. 7.8) for a discussion of denotation failure. Or in the additional literature cited in the section 3.1. You can take an argument from this literature and carefully discuss it (e.g., find objections or evaluate its potentially hidden assumptions). But, even better, you can also come up with your own arguments.

Exercise 3.b (Practice)**.** A consequence relation $\vDash$ is said to have the *deduction theorem* if it satisfies

$$\Gamma, \varphi \vDash \psi \text{ iff } \Gamma \vDash \varphi \to \psi.$$

For this exercise you can, in particular, assume the statements of exercise 3.7 and 3.8.

In other words, the metalanguage inference $\vDash$ can be expressed by the object language inference $\to$.

(a) Show that classical logic $\vDash_{\mathsf{CL}}$ has the deduction theorem.

(b) Show that none of the logics $\mathsf{K}_3^{\mathsf{s}}, \mathsf{Ł}_3, \mathsf{LP}$ has the deduction theorem. Hint: Take $\Gamma = \emptyset$ and consider $(p \wedge \neg p) \to q$.

So, by exercise 3.8, also ST *has the deduction theorem.*

Exercise 3.c (Practice/problem)**.** Do exercises 3.7 and 3.8.

Exercise 3.d (Problem)**.** This exercise provides another explanation of the truth-tables for FDE by separating truth-making ($\vDash^+$) from false-making ($\vDash^-$). (We'll only consider the connectives $\neg, \wedge, \vee$ in this exercise.) Similarly to thinking of classical valuations as classical states, think of a valuation $v : \mathsf{P} \to \{0, 1, \mathfrak{b}, \mathfrak{n}\}$ as a state, but now one that needn't be complete nor consistent: so besides making an atomic proposition p true (and not false) or false (and not true), it can also make p both true and false (inconsistent) or it can make p neither true nor false (incomplete). Thus, being true is independent of being false and vice versa (while in classical logic, one determines the other). We can write, e.g., $v \vDash^+ p$ and $v \nvDash^- p$ to say that v makes true p but it doesn't make false p (so $v(p) = 1$). More generally, we

As in the classical state case, we could separate what states are from what they do: i.e., call any object a state as long as it determines a valuation $v : \mathsf{P} \to \{0, 1, \mathfrak{b}, \mathfrak{n}\}$. Thus, e.g., a database could be regarded as a state.

define truth-making and false-making by induction:

$$\begin{aligned}
\nu \vDash^+ p &\Leftrightarrow \nu(p) \in \{1, b\} & \nu \vDash^- p &\Leftrightarrow \nu(p) \in \{0, b\} \\
\nu \vDash^+ \neg\varphi &\Leftrightarrow \nu \vDash^- \varphi & \nu \vDash^- \neg\varphi &\Leftrightarrow \nu \vDash^+ \varphi \\
\nu \vDash^+ \varphi \wedge \psi &\Leftrightarrow \nu \vDash^+ \varphi \text{ and } \nu \vDash^+ \psi & \nu \vDash^- \varphi \wedge \psi &\Leftrightarrow \nu \vDash^- \varphi \text{ or } \nu \vDash^- \psi \\
\nu \vDash^+ \varphi \vee \psi &\Leftrightarrow \nu \vDash^+ \varphi \text{ or } \nu \vDash^+ \psi & \nu \vDash^- \varphi \vee \psi &\Leftrightarrow \nu \vDash^- \varphi \text{ and } \nu \vDash^- \psi
\end{aligned}$$

(a) As motivated, on propositional atoms the four truth-values $\{0, 1, b, n\}$ correspond to the four possible combinations of (not) truth-making and (not) false-making. Show that this extends to all formulas, i.e., show, by induction on φ, that

Here you take the FDE *truth-tables as given and 're-express' them in terms of truth-making and false-making.*

$$\begin{aligned}
\nu(\varphi) = 1 &\Leftrightarrow \nu \vDash^+ \varphi \text{ and } \nu \nvDash^- \varphi \\
\nu(\varphi) = b &\Leftrightarrow \nu \vDash^+ \varphi \text{ and } \nu \vDash^- \varphi \\
\nu(\varphi) = n &\Leftrightarrow \nu \nvDash^+ \varphi \text{ and } \nu \nvDash^- \varphi \\
\nu(\varphi) = 0 &\Leftrightarrow \nu \nvDash^+ \varphi \text{ and } \nu \vDash^- \varphi
\end{aligned}$$

But, since doing all cases is too much work, only show for each induction case (atomic, negation, conjunction, disjunction) one of the equivalences (you can choose which; some are harder and some easier).

(b) This explains why $b \wedge n = 0$: if $\nu(p) = b$ and $\nu(q) = n$, then $\nu \vDash^+ p, \nu \vDash^- p, \nu \nvDash^+ q, \nu \nvDash^- q$, so $\nu \nvDash^+ p \wedge q$ (since $\nu \nvDash^+ q$) and $\nu \vDash^- p \wedge q$ (since $\nu \vDash^- p$), so $\nu(p \wedge q) = 0$. Give a similar explanation for $\neg b = b$ and $b \vee n = 1$.

Here you explain some entries of the FDE *truth-tables in the (hopefully natural) terms of truth-making and false-making.*

Exercise 3.e (Philosophical)**.** In this exercise, you should discuss the problems for many-valued logics posed by higher-order vagueness. You can pick a particular logic and a particular aspect of higher-order vagueness that you find interesting and discuss it in greater detail by developing your own ideas. You might ask: Does higher-order vagueness even pose a problem? If so, do many-valued logics even need to respond to it? If so, can they? For example, can the problems with a three-valued definition of Δ be explained away? You can find inspiration in the quoted literature. If you prefer a more concrete question, discuss whether the solution to the sorites paradox suggested by ST can deal with higher-order vagueness (for inspiration see Cobreros et al. (2015, sec. 4.2)).

Exercise 3.f (Problem)**.** In 1952, Kleene introduced the weak and strong three-valued logics in the influential textbook 'Introduction to Metamathematics' (Kleene 1952, p. 334 f.). Kleene discusses there a sense in which the two logics K_3^w and K_3^s are weak and strong, respectively. Can you verify

the following claim made there? Kleene calls a truth-table *regular* if

> "A given column (row) contains 1 in the i row (column), only if the column (row) consists entirely of 1's; and likewise for 0" (p. 334; adjusted notation: using our 1, 0, i instead of Kleenes t, f, u, respectively).

And the claim is

> "[The] strong tables are uniquely determined as the strongest possible regular extensions of the classical 2-valued tables, i.e., they are regular, and have a 1 or [a] 0 in each position where any regular extension of the 2-valued tables can have a 1 or [a] 0 (whether 1 or 0 being uniquely determined)" (p. 335; notation again adjusted).

Can you similarly describe a sense in which the weak tables are weak?

3.5 Notes

Partly based on Sider (2010, sec. 3.3–3.4) and Priest (2008, ch. 7–8).

4 Infinitely-valued logic

So far, we've only seen logics characterized using only finitely many truth-values. In the algebraic semantics, we've already seen that we *can* also use infinitely many truth-values (using Boolean algebras for the semantics of classical logic). But so far this could equivalently be done with finitely many. In this chapter, we consider two examples of logics using infinitely many truth-values.

The first, fuzzy logic, builds in infinitely many truth-values from the start motivated by vague concepts. The second, intuitionistic logic, was not motivated to use infinitely many truth-values but rather to capture a verificationalist point of view—formalized by a state-based semantics (intuitionistic Kripke models). However, concerning the question of whether we can also formalize intuitionistic logic with a truth-value semantics, we prove Gödel's result showing that intuitionistic logic needs infinitely many truth-values. This algebraic semantics is then given using Heyting algebras (the intuitionistic counterparts to Boolean algebras). We show that the state-based and algebraic semantics are equivalent, and we use the advantage of the algebraic semantics in providing a simple completeness proof.

Key concepts
- Fuzzy logic, continuously many truth-values, generalization of $Ł_3$.
- Intuitionism in mathematics (vs. Platonism), truth as provability (or verifiability), BHK-interpretation.
- State-based semantics: intuitionistic Kripke models.
- Gödel's theorem: intuitionistic logic is not many-valued.
- Algebraic semantics: Heyting algebras.
- Equivalence of state-based and algebraic semantics.
- Algebraic completeness proof: Lindenbaum–Tarski algebra.
- Relation classical and intuitionistic logic: Glivenko's theorem.
- Disjunction property.

4.1 Fuzzy logic

4.1.1 Motivation

The standard motivation for fuzzy logic is vagueness: We've already seen that, if we have a sorites sequence (with N large)

$$\begin{aligned} p_1 &= 1 \text{ grain of sand makes a heap} \\ &\vdots \\ p_N &= N \text{ grains of sand make a heap} \end{aligned}$$

we'd intuitively say that the first few p_j's are definitely false, the last few p_j's are definitely true, but the p_j's in between are undetermined.

The idea of many-valued logics was to introduce a new truth-value i standing for undetermined—and take the in-between p_j's to have truth-value i. The resulting logic is naturally given by either K_3^s (strong Kleene) or Ł$_3$ (Łukasiewicz). That's what we've seen in the last chapter. But there we've also seen an objection (section 3.3.3). A clear-cut border between the definite case (1 or 0) and the borderline case (i) is just as problematic as a clear-cut border between 1 and 0. We wouldn't know where that border should lie just as much as we wouldn't know where a border between 1 and 0 should lie.

Once accepting one non-classical truth-value, a natural continuation of the idea is to allow (infinitely) many: namely continuously many truth-values between 0 and 1 to 'even out' any discrete boundary between truth-values. Fuzzy logic describes the logic(s) on this choice of truth-values.

4.1.2 Formal logic

Following the truth-value semantics template, the arguably most common fuzzy logic Ł$_c$ is described as follows. The notation Ł$_c$ is to indicate that this logic is the Łukasiewicz logic Ł$_3$ extended to continuously many truth-values.

There are other ones, described via so-called t*-norms (Priest 2008, sec. 11.7a), but this is all we need here.*

- The set of truth-values T is the unit interval $[0,1] = \{x \in \mathbb{R} : 0 \leqslant x \leqslant 1\}$ (i.e., all real numbers between and including 0 and 1).
- A fuzzy valuation is a function $v : P \to [0,1]$ that extends to all formulas by interpreting the connectives with the following functions

on $[0,1]$:

$$\begin{aligned}
\neg x &:= 1 - x \\
x \wedge y &:= \min(x, y) \\
x \vee y &:= \max(x, y) \\
x \to y := x \ominus y &:= \begin{cases} 1 & \text{if } x \leqslant y \\ 1 - (x - y) & \text{if } x > y \end{cases} \\
\bot &:= 0 \\
\top &:= 1
\end{aligned}$$

(We'll shortly motivate this choice.)

- Consequence $\Gamma \models_{Ł_c} \varphi$ is defined as: For any $0 \leqslant \epsilon \leqslant 1$ and for any fuzzy valuation $v : \mathsf{P} \to [0,1]$, if $v(\psi) \geqslant \epsilon$ for all $\psi \in \Gamma$, then $v(\varphi) \geqslant \epsilon$. (Again, we'll shortly motivate this choice.)

The choice of truth-values $\mathsf{T} = [0,1]$ is motivated by vagueness, as seen above. But what about the truth-functions? The Boolean connectives $\neg, \wedge, \vee, \bot, \top$ are the literal generalization of the classical truth-functions: now applied to any $x, y \in [0,1]$ rather than just $x, y \in \{0,1\}$. So the conditional $\to$ is the one needing explanation: If $x \leqslant y$, then x is less true than (or equally true as) y, so $x \to y = 1$ as in classical logic. But if $x > y$, we move, in the conditional, to something less true, so there is something wrong with that 'inference'. Hence $x \to y$ shouldn't be 1; in fact, it is just as much less than 1 as we 'lost' in the inference from x to y, i.e., $x \to y = 1 - (x - y)$.

The consequence relation can be understood as follows. In any context where we apply fuzzy valuations (e.g., a context involving a vague predicate), we have a standard of what we'd take as acceptable. For example, if I want a red apple, it is acceptable for me if I get a red apple with tiny patch of green (so 'this apple is red' is, say, 0.95 true). In this context, 'acceptable' means $\geqslant 0.95$. So, generally speaking, a context determines an acceptability parameter ϵ (with $0 \leqslant \epsilon \leqslant 1$), and 'acceptable' means $\geqslant \epsilon$. Thus, in a context ϵ, we want our reasoning to be acceptability preserving: whenever the premises are acceptable (i.e., get value $\geqslant \epsilon$), also the conclusion is acceptable. However, we don't want our logic to be context-sensitive. (A generally important feature of logic precisely is that it is 'universal', i.e., context independent.) So we speak of consequence if we have acceptability-preservation in *any* context. This is precisely the

above definition of $\models$.

We end this subsection with several results about this formal logic. First, modus ponens fails:

Example 4.1. We have $p, p \to q \not\models_{Ł_c} q$: Indeed, take $\epsilon := 0.9$ and a fuzzy valuation ν with $\nu(p) = 0.9$ and $\nu(q) = 0.8$. Then $\nu(p) = 0.9 \geqslant \epsilon$ and $\nu(p \to q) = 1 - (\nu(p) - \nu(q)) = 1 - (0.1) = 0.9 \geqslant \epsilon$, but $\nu(q) = 0.8 \ngeqslant \epsilon$. ⌟

The consequence relation $\models_{Ł_c}$ has a simple characterization:

If $X \subseteq \mathbb{R}$ is a nonempty set of real numbers, $\inf X$ *is the greatest real number smaller than every element of X.*

Exercise 4.2. Show (for $\Gamma \neq \emptyset$): $\Gamma \models_{Ł_c} \varphi$ iff for all fuzzy valuations ν, $\inf\{\nu(\psi) : \psi \in \Gamma\} \leqslant \nu(\varphi)$.

A version of $Ł_c$, which takes $\epsilon = 1$ as the only designated value, is known as the *Łukasiewicz continuum-valued logic* $Ł_\aleph$. They are related as follows:

The Hebrew letter $\aleph$ (aleph) is used in set theory to denote cardinalities (sizes of sets): $\aleph_1$ is the first uncountable cardinal number. Under the continuum hypothesis, it is the size of the unit interval $[0, 1]$.

Exercise 4.3. Show: $\psi_1, \ldots, \psi_n \models_{Ł_c} \varphi$ iff $\models_{Ł_\aleph} \psi_1 \wedge \ldots \wedge \psi_n \to \varphi$.

The fuzzy logic $Ł_c$ generalizes both classical logic and Łukasiewicz logic in the following sense:

Exercise 4.4.

1. Call a fuzzy valuation $\nu : \mathsf{P} \to [0, 1]$ *classical* if $\nu(p) \in \{0, 1\}$ for each $p \in \mathsf{P}$. Show that, for such ν, we have, for any $\varphi \in \mathcal{L}_{\text{prop}}$, that $\nu_{Ł_c}(\varphi) = \nu_{CL}(\varphi)$. Conclude: if $\Gamma \models_{Ł_c} \varphi$, then $\Gamma \models_{CL} \varphi$.
2. Call a fuzzy valuation $\nu : \mathsf{P} \to [0, 1]$ *three-valued* if $\nu(p) \in \{0, \frac{1}{2}, 1\}$ for each $p \in \mathsf{P}$. We can think of such ν as a valuation $\mathsf{P} \to \{0, 1, i\}$ by identifying $i = \frac{1}{2}$. Given this identification, show that, for such ν we have, for any $\varphi \in \mathcal{L}_{\text{prop}}$, that $\nu_{Ł_c}(\varphi) = \nu_{Ł_3}(\varphi)$. Conclude: if $\Gamma \models_{Ł_c} \varphi$, then $\Gamma \models_{Ł_3} \varphi$.

4.1.3 Assessment

We'll be brief here. One issue is the failure of modus ponens: As we've seen, the fuzzy logic $Ł_c$ generalizes the three-valued Łukasiewicz logic $Ł_3$. However, the latter satisfies modus ponens (check!). So, in a sense, giving up modus ponens is the price to pay when moving from the discrete three-valued approach to vagueness (definitely true, definitely false, indefinite) to the continuous fuzzy-valued approach. Two attempts to alleviate these worries are: First, modus ponens fails only once we move to fuzzy values; for crisp classical values it still holds. Second, maybe we should expect modus ponens to fail if we want to avoid the sorites paradox: after all, the premises are plausible, so as soon as we also allow all instances of modus ponens we can derive the implausible conclusion.

Exercise 4.5. Think how this last comment reconciles with how ST-logic keeps modus ponens and still offers a solution to the sorites paradox (as discussed in section 3.3.1).

Exercise 4.6. In exercise 4.4 (2) you have shown that fuzzy logic generalizes the three-valued Łukasiewicz logic. Is giving up modus ponens an appropriate price to pay for the higher expressive power of describing sorites sequences?

Another issue to consider is the following.

Exercise 4.7. Think about the following objection: the move in a sorites sequence from 1 to any value $\neq 1$ is a substantial qualitative change, which is as problematic as a cut-off between 1 and $\mathfrak{i}$.

4.2 Intuitionistic logic

4.2.1 Motivation

Truth is a complicated concept: Even without saying what truth *is*, but only making modest assumptions about what it *does*, we get the liar paradox—as we've already seen. If we look at what truth *is*, a typical explanation of why a sentence like 'Snow is white' is true is that it is in *correspondence* with a fact about the world: namely, that the real stuff out there which is snow has the property of having the color white, i.e., reflecting light of the appropriate wave-lengths. This is known as as the *correspondence theory of truth* (David 2020). At least for such everyday sentences this seems to be pretty close to our intuitions: that there is a real world independent of us and we can say things about it—some true, others false.

A classic example inspired by the seminal paper of Tarski (1956) who used 'it is snowing'.

But it gets more problematic for sentences involving more abstract concepts: as in mathematics. For example, why is the sentence 'There are infinitely many prime numbers' true? According to the correspondence theory the explanation would be something like: there exist objects—the natural numbers—some of which have the objective property of being prime, and there in fact are infinitely many that have it. This position that mathematical objects—like numbers, functions, sets, etc.—exist as independent objects having objective properties in a universe of abstract objects is known as *Platonism* in the philosophy of mathematics (in reminiscence of Platonic ideas).

The classic proof of the infinitude of primes is Euclid's (~300 BC). A favorite of mine is Furstenberg's topological proof.

However, not everyone is comfortable with accepting such a strong metaphysical claim. A rivaling view is *intuitionism* introduced by the

The third of the three prominent views at the time was formalism*: mathematics as deriving sentences from some axioms according to some rules.*

Dutch mathematician L.E.J. Brouwer during the foundational crisis of mathematics at the end of the 19th century and the beginning of the 20th century.

Rather than taking mathematical truths to be *discoveries* about the Platonic universe, intuitionism takes them to be *creations* of the mind: Mathematical objects are constructed by the mathematician, and 'truths' are statements about these objects for which the mathematician has a proof. After all, if an external world is not available to describe truth, an internal understanding of truth as provability or verifiability is natural.

This doesn't mean that 'anything goes': the constructions still have to follow 'rational thought'. Much like constructions in the real world—like bridges, cars, etc.—also have to follow physical laws.

This shift in perspective on 'truth' has a dramatic effect on the associated logic: While classical logic describes preservation of an objective external truth, *intuitionistic logic* describes preservation of verifiable internal truth—i.e., preservation of constructability and provability. In this section, we work out precisely what this intuitionistic logic has to look like. But let's start with a concrete example why it has to differ.

In classical logic, the law of excluded middle, $\varphi \vee \neg\varphi$ is valid, but we shouldn't expect it to be valid intuitionistically: we may be in a situation where we neither have a proof of φ nor a proof of $\neg\varphi$. So φ is like a mathematical conjecture that has neither been proven nor disproven. Concretely, this shows up in the following standard example:

See e.g. OpenLogicProject (2021, ch. 55, p.761 f.)

> *Theorem.* There are irrational numbers x and y such that x^y is rational.
>
> *Proof.* Famously, $\sqrt{2}$ is an irrational number. We consider $z := \sqrt{2}^{\sqrt{2}}$. Either z is rational or irrational. If z is rational, choose $x := y := \sqrt{2}$, whence x and y are irrational but $x^y = z$ is rational. If z is irrational, then $x := z$ and $y := \sqrt{2}$ are irrational, but $x^y = \sqrt{2}^{\sqrt{2}\sqrt{2}} = \sqrt{2}^2 = 2$ is rational.

Classically, the proof is fine, but intuitionistically it is not since it uses the law of excluded middle (in the third sentence). And, indeed, it did not preserve constructability: it didn't concretely construct two irrational numbers x and y such that x^y is rational—it merely showed there *have to* be such numbers without actually constructing them.

So, given that provability/verifiability provides another example for indeterminacy (cf. section 3.1.2), we might try to describe intuitionistic logic as a three-valued logic: either we have a proof for φ (so $\nu(\varphi) = 1$) or we have a disproof for φ, i.e., a proof for $\neg\varphi$ (so $\nu(\varphi) = 0$) or we have neither nor (so $\nu(\varphi) = i$). However, the typical three-valued logics satisfy double negation elimination: $\neg\neg\varphi \vDash \varphi$. But this also doesn't preserve provability: roughly speaking, showing that we cannot disprove φ doesn't

provide an explicit proof of φ. Below we'll show that, in fact, no finitely-valued logic can describe intuitionistic logic.

So how then develop intuitionistic logic more systematically? In classical logic, we used truth-conditions to describe the meaning of formulas (and connectives). In intuitionistic logic, the meaning of a formula is described by the proofs that it has—this is known as the *BHK-interpretation* (for Brouwer, Heyting and Kolmogorov). Specifically, in place of the truth-conditions, we have:

- A proof of $\varphi \wedge \psi$ consists of a proof of φ and a proof of ψ.
- A proof of $\varphi \vee \psi$ consists of a proof of φ or of a proof of ψ.
- A proof of $\neg\varphi$ consists of a proof that there is no proof of φ.
- A proof of $\varphi \to \psi$ consists of a method of converting any proof of φ to a proof of ψ.

Note: This is *not* a formal semantics, rather it is an informal description of how to interpret formulas intuitionistically. In particular, 'proof' is understood intuitively (as a convincing mathematical argument), not as a formal derivation in a proof system.

4.2.2 Formal logic

State-based semantics (intuitionistic Kripke models)

In 1965, Saul Kripke provided a state-based semantics for intuitionistic logic that aims to formalize the above intuition of "constructing mathematical truths over time". Following the template of a state-based semantics (see section 2.3.1), it is defined as follows. An intuitive motivation is given below.

Intuitionistic logic itself was formulated by Heyting—a student of Brouwer—in 1930, and conncetions to modal logic have already been established by Gödel in 1933.

Definition 4.8. First, an *intuitionistic Kripke model* M is a triple (S, R, I) where

- S is a nonempty set (the state space),
- $R \subseteq S \times S$ is a reflexive and transitive relation, and
- $I : S \times P \to \{0, 1\}$ is a function (the interpretation function, so $I(s, p) = 1$ means p is true at state s).

Reflexive means: sRs *(for all* $s \in S$*). Transitive means: if* s_1Rs_2 *and* s_2Rs_3*, then* s_1Rs_3 *(for all* $s_1, s_2, s_3 \in S$*). Aka persistence*

such that the following so-called *heredity condition* is satisfied

$(*)$ for all $s, s' \in S$, if $I(s, p) = 1$ and sRs', then $I(s', p) = 1$.

Second, by recursion on formulas, we define the notion of a state making true a formula: $M, s \models \varphi$.

- $M, s \models p$ iff $I(s, p) = 1$
- $M, s \models \top$ always and $M, s \models \bot$ never
- $M, s \models \varphi \wedge \psi$ iff $M, s \models \varphi$ and $M, s \models \psi$
- $M, s \models \varphi \vee \psi$ iff $M, s \models \varphi$ or $M, s \models \psi$
- $M, s \models \varphi \to \psi$ iff, for all $s' \in S$, if sRs' and $M, s' \models \varphi$, then $M, s' \models \psi$.
- $M, s \models \neg\varphi$ iff for all $s' \in S$, if sRs', then $M, s' \not\models \varphi$.

  Equivalently, $M, s \models \varphi \to \bot$.
- $M, s \models \varphi \leftrightarrow \psi$ iff $M, s \models \varphi \to \psi$ and $M, s \models \psi \to \varphi$.

Often, $\neg$ and $\leftrightarrow$ are not considered to be part of the intuitionistic language, but only as abbreviations: $\neg\varphi := \varphi \to \bot$ and $\varphi \leftrightarrow \psi := (\varphi \to \psi) \wedge (\psi \to \varphi)$.

Third, consequence is defined as: $\Gamma \models_{\mathsf{IL}} \varphi$ iff, for all models $M = (S, R, I)$ and states $s \in S$, if $M, s \models \psi$ for all $\psi \in \Gamma$, then $M, s \models \varphi$.

Before getting to the motivation, we first show that the heredity condition extends to all formulas:

Proposition 4.9. *For all formulas φ, all models $M = (S, R, I)$, and all states $s, s' \in S$: if $M, s \models \varphi$ and sRs', then $M, s' \models \varphi$.*

Proof sketch. By induction on φ: For atomic φ, this is the heredity condition.

For $\varphi = \psi \wedge \chi$, if $M, s \models \psi \wedge \chi$ and sRs', then $M, s \models \psi$ and $M, s \models \chi$, so the induction hypothesis implies $M, s' \models \psi$ and $M, s' \models \chi$, so $M, s' \models \psi \wedge \chi$.

We skip the other connectives and only look at $\varphi = \psi \to \chi$. Assume $M, s \models \psi \to \chi$ and sRs', and show $M, s' \models \psi \to \chi$. So let $s'' \in S$ with $s'Rs''$ and $s'' \models \psi$, and show $s'' \models \chi$. By transitivity, sRs''. So, by definition of $s \models \psi \to \chi$, $s'' \models \psi$ implies $s'' \models \chi$. □

Now, the motivation of the intuitionistic Kripke semantics is this: The states of a model are the possible states of mind of a possible (ideal) mathematician. The relation R describes the possible future states that the mathematician can reach (next). Thus, a formula φ is true at a state if the mathematician has a proof for φ. Being ideal, they don't forget proofs, so φ is true in all future states—which is the (extended) heredity condition. A disproof is modeled as 'false in all future states', so the mathematician knows there is no proof. A branching in R with a φ-branch

It's worth reflecting on how well this captures the original motivation for intuitionistic logic.

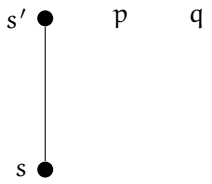

Figure 4.1: A Kripke model showing that the intuitionistic conditional is not the material conditional.

and a $\neg\varphi$-branch thus indicates that the mathematician is still undecided about φ.

To give a concrete example of a Kripke model, we construct one showing that the intuitionistic conditional is *not* the material conditional.

Example 4.10. Consider the Kripke model depicted in figure 4.1: Its state space is $S = \{s, s'\}$. The relation R is such that s is 'less than' s', i.e., $R = \{(s, s), (s, s'), (s', s')\}$. Then interpretation I is such that p and q are both false at s (so they are not written next to s), but they are both true at s' (so they are written next to s'). All other propositional atoms are set to false at every state. Thus, R clearly is transitive. And I satisfies the heredity condition: as you 'go up' along the 'order' R, sentences only turn from false to true, but never the other way round.

Now, let's see that $s \vDash p \to q$. Assume sRs'' and $s'' \vDash p$, and show $s'' \vDash q$. By definition of R, the only options are $s'' = s$ or $s'' = s'$. But $s'' = s$ cannot be since $s \nvDash p$ but $s'' \vDash p$. So $s'' = s'$. Then $s'' = s' \vDash q$.

Let's see that $s \nvDash \neg p \vee q$, i.e., $s \nvDash \neg p$ and $s \nvDash q$. The latter holds by construction. For the former, note that sRs' and $s' \vDash p$, so, by definition, $s \nvDash \neg p$.

Hence, our Kripke model witnesses that $\varphi \to \psi \nvDash_{IL} \neg\varphi \vee \psi$. ⌟

Intuitionistic logic is not many-valued

As a state-based semantics, the intuitionistic Kripke semantics cannot deliver the other, algebraic intuition: that formulas have truth-values and that their meanings—i.e., propositions—have algebraic structure. So we turn to the question whether this can be done. We start looking at truth-values and find that finitely many won't do.

We've seen that the typical three-valued logics fail to provide a semantics to intuitionistic logic (by taking the third truth-value i as neither provable

nor disprovable). Much deeper, a result of Gödel shows that in fact any approach using finitely many truth-values is doomed to fail.

Theorem 4.11 (Gödel (1932))**.** *Intuitionistic logic cannot be viewed as a many-valued logic: There is* no *finite set* T *of truth-values and a subset* $\mathsf{D} \subseteq \mathsf{T}$ *of designated values, together with an interpretation of* $\neg, \wedge, \vee, \to, \leftrightarrow, \bot, \top$ *as truth-functions (of respective arity) on* T*, such that* $\Gamma \models_{\mathsf{IL}} \varphi$ *iff for all valuations* $\nu : \mathsf{P} \to \mathsf{T}$*, if* $\nu(\psi) \in \mathsf{D}$ *for all* $\psi \in \Gamma$*, then* $\nu(\varphi) \in \mathsf{D}$*.*

Gödel's result appeared shortly after his incompleteness theorems. This led Heyting to respond: "It is as if you had a malicious pleasure in showing the purposelessness of others' investigations" (see van Atten 2017, sec. 4.5.1).

The proof is established, step by step, in exercise 4.c.

Three further remarks on the result (just to put it into perspective, but we won't cover them here): First, the proof also shows that there are infinitely many logics between classical logic and intuitionistic logic—they are known as intermediate logics. Second, a similar proof also works for a lot of modal logics, i.e., showing that they, too, cannot have a semantics with finitely many truth-values. Third, the deeper connection behind this is that there are translations from intuitionistic/intermediate logics to modal logics (Gödel–McKinsey–Tarski translation). The modal logic thus corresponding to an intermediate logic is known as its (or rather a) modal companion.

Algebraic semantics (Heyting algebras)

So intuitionistic logic is *not* finitely-valued. But we don't know yet that it *is* infinitely-valued. And we don't know yet the algebraic structure of intuitionistic propositions. We need this to respond to Gödel's impossibility result (theorem 4.11) with a possibility result: that if we allow infinitely many truth-values, we still can given a truth-value/algebraic semantics to intuitionistic logic.

Recall that Boolean algebras turned out to provide a good notion of truth-value and proposition for classical logic. For intuitionistic logic, the appropriate notion of an algebra is that of a so-called Heyting algebra. Again, this will be a 'functionalist' and not 'ontological' description of truth-values/propositions: saying how they behave (i.e., what their algebraic structure is), rather than what they are.

Definition 4.12. A *Heyting algebra* is a structure $(A, \vee, \wedge, \to, 0, 1)$ where $\vee, \wedge, \to \colon A \times A \to A$ and $0, 1 \in A$ such that

1. $(A, \vee, \wedge, 0, 1)$ is a bounded distributive lattice, i.e., it satisfies axioms 1 (lattice), 2 (distributive), and 3 (bounded by 0 and 1) of definition 2.11.

2. $a \to b$ is the greatest element $c \in A$ such that $a \wedge c \leqslant b$.

The operation $\neg x$ is defined as $x \to 0$ and hence is not explicitly mentioned in the signature (i.e., in the list of operations $\vee, \wedge, \to, 0, 1$). Again, we sometimes just write A to refer to the Heyting algebra $(A, \vee, \wedge, \to, 0, 1)$.

We soon explain this definition, but first: The algebraic semantics for intuitionistic logic then accordingly is as follows:

Definition 4.13 (Heyting algebraic semantics)**.** Let HA be the class of Heyting algebras. As in the template, if $A = (A, \vee, \wedge, \to, 0, 1)$ is in HA, an A-valuation is a function $v : \mathsf{P} \to A$ extended recursively to all formulas by

$$\begin{aligned} v(\varphi \vee \psi) &= v(\varphi) \vee v(\psi) & v(\bot) &= 0 \\ v(\varphi \wedge \psi) &= v(\varphi) \wedge v(\psi) & v(\top) &= 1 \\ v(\neg\varphi) &= \neg v(\varphi) = v(\varphi) \to 0 & v(\varphi \to \psi) &= v(\varphi) \to v(\psi) \end{aligned}$$

and $v(\varphi \leftrightarrow \psi) = \big(v(\varphi) \to v(\psi)\big) \wedge \big(v(\psi) \to v(\varphi)\big)$.

Consequence $\Gamma \models_{\mathsf{IL}} \varphi$ according to the *Heyting algebraic semantics* is defined as

> For all $A \in \mathsf{HA}$ and A-valuations v, if $v(\psi) = 1$ for all $\psi \in \Gamma$, then $v(\varphi) = 1$.

For emphasis, we also write $\Gamma \models^{\mathsf{A}}_{\mathsf{IL}} \varphi$ for the algebraic consequence relation and $\Gamma \models^{\mathsf{S}}_{\mathsf{IL}} \varphi$ for the state-based one, but we'll soon show that they are identical.

In the rest of this subsection, we explain these definitions (mostly that of a Heyting algebra).

First, how to interpret axiom 2? To start, assume $a \leqslant b$. Then the move from a to b is 'truth-preserving', i.e., we move from a to something that is more (or equally) true, so the conditional $a \to b$ should be true. Indeed: then $a \wedge 1 = a \leqslant b$, so, since 1 already is the largest element of A, it in particularly is the largest element c such that $a \wedge c \leqslant b$.

But now, if $a \not\leqslant b$, then the move from a to b is not 'truth-preserving', so the conditional $a \to b$ shouldn't be 1. But how much less than 1 should it be? The axiom says: Consider the additional assumptions c such that the move from a together with c to b is still truth-preserving (i.e., $a \wedge c \leqslant b$). The closer c is to 1, the less is required to make the conditional $a \to b$ truth-preserving, so the closer $a \to b$ is to being true. Thus, the largest

Notice the similarity to the discussion of the conditional in fuzzy logic.

such c is the minimal assumption needed to make $a \to b$ truth-preserving. Hence this c describes how close to true $a \to b$ is, and hence serves as the truth-value of $a \to b$—and axiom 2 ensures the existence of this c.

Second, another way to put this is to think of the elements of A again as propositions. If c is such that $a \wedge c \leqslant b$, then c is a propositions for which modus ponens is sound (with respect to a and b): moving from a and c to b is truth-preserving. So $a \to b$ is the weakest sentence (i.e., 'most true' or 'least contestable to assume') for which modus ponens is sound. Formally, this can be expressed as follows:

Such kinds of equivalences may be familiar if you've seen Galois connections (in order theory) or adjunctions (in category theory). This leads to the general concept of residuated lattices.

Exercise 4.14. If A is a Heyting algebra and $a, b \in A$, then, for any $c \in A$,

$$a \wedge c \leqslant b \text{ iff } c \leqslant a \to b.$$

Third, strictly speaking, axiom 2 is not an equation (as it should be for a proper 'algebraic' definition). But one can show that one can equivalently state it as the (conjunction of the) following two axioms

See e.g. Bezhanishvili and de Jongh (2006, Thm. 50).

- $(x \to y) \wedge y = y$ and $x \wedge (x \to y) = x \wedge y$.
- $x \to (y \wedge z) = (x \to y) \wedge (x \to z)$ and $(x \vee y) \to z = (x \to z) \wedge (y \to z)$.

Fourth, Heyting algebras are a generalization of Boolean algebras:

Think about why this is to be expected.

Exercise 4.15. Show that any Boolean algebra $(A, \vee, \wedge, \neg, 0, 1)$ can be regarded as a Heyting algebra $(A, \vee, \wedge, \to, 0, 1)$ with $x \to y := \neg x \vee y$. (Recall that, in classical logic, the conditional $\to$ is the material conditional.)

Fifth, a typical example of Heyting algebras that aren't Boolean algebras are finite chains (of length $\geqslant 3$):

Exercise 4.16. Let $S_n = \{s_1, \ldots, s_n\}$ be a linear order of n elements (i.e., $s_i \leqslant s_j$ iff $i \leqslant j$). Define

- $s_i \vee s_j := \max(s_i, s_j)$ and $s_i \wedge s_j := \min(s_i, s_j)$
- $s_i \to s_j := \max(s_k : s_i \wedge s_k \leqslant s_j)$
- $0 := s_1$ and $1 := s_n$.

Show that $(S_n, \vee, \wedge, \to, \bot, \top)$ is a Heyting algebra: i.e., show that the above functions are well-defined and satisfy axioms 1 and 2.

Sixth, in particular, the linear order $S_3 = \{0, i, 1\}$ (with $0 \leqslant i \leqslant 1$) is the smallest Heyting algebra that is not a Boolean algebra. We know the set $\{0, i, 1\}$ as the set of truth-values from three-valued logics. How does intuitionistic logic differ here?

Exercise 4.17. The truth-functions for $\neg, \wedge, \vee, \to$ over the set of truth-values S_3 provided by the Heyting algebra operations are the following.

$\neg$	
1	0
i	0
0	1

$\wedge$	1	i	0
1	1	i	0
i	i	i	0
0	0	0	0

$\vee$	1	i	0
1	1	1	1
i	1	i	i
0	1	i	0

$\to$	1	i	0
1	1	i	0
i	1	1	0
0	1	1	1

Pick a few entries of the truth-tables, cover them, and make sure you calculate the same value for them yourself. So the truth-function for $\wedge$ and $\vee$ are the same as in strong Kleene logic K_3^s (and also $Ł_3, \mathsf{LP}, \mathsf{ST}$). But those for $\neg$ and $\to$ differ.

Seventh, from Gödel's theorem (theorem 4.11) we know that intuitionistic logic $\models_{\mathsf{IL}}$ cannot be described as a finitely-valued logic. In particular, the truth-functions provided by the Heyting algebra S_3 cannot describe all of intuitionistic logic. In fact, no S_n can, as the following exercise shows (so it establishes a special case of Gödel's theorem):

Exercise 4.18.
1. Show that the sentence $(\varphi \to \psi) \vee (\psi \to \varphi)$ is valid on any Heyting algebra S_n: i.e., given n, show that, for any S_n-valuation v, $v\big((\varphi \to \psi) \vee (\psi \to \varphi)\big) = 1$.
2. Show that $\not\models_{\mathsf{IL}} (\varphi \to \psi) \vee (\psi \to \varphi)$, either using Kripke models or Heyting algebras.

When adding this sentence as an axiom to intuitionistic logic one obtains the so-called Gödel–Dummett logic.

Eighth, if you know the concept of a topological space (X, τ), another example of a Heyting algebra is the set of open sets τ ordered by inclusion: so $\vee$ is union, $\wedge$ is intersection, $0 = \emptyset$, $1 = X$, and $U \to V$ is defined as the topological interior of $U^c \cup V$ (i.e., the largest open set contained in the 'material conditional' $U^c \cup V$). One can show that any Heyting algebra is a subalgebra of the open sets of some topological space.

A topological space (X, τ) *is a set* X *together with a collection* τ *of subsets of* X *closed under arbitrary unions and finite intersection with* $\emptyset \in \tau$. *The elements of* τ *are called* open sets.

Intuitionistic logic is infinitely-valued: duality

Now, we've given an algebraic semantics which can use infinitely many truth-values. To show that intuitionistic logic is infinitely-valued, we still have to show that this algebraic semantics actually is a semantics for intuitionistic logic (as given by the Kripke semantics), i.e., that $\models^{\mathsf{A}}_{\mathsf{IL}} = \models^{\mathsf{S}}_{\mathsf{IL}}$. In other words: we show that, as in classical logic, the state-based and algebraic semantics coincide. We'll only deal with validity: that's easier and still contains the main ideas.

Theorem 4.19. $\models^{\mathsf{A}}_{\mathsf{IL}} \varphi$ *iff* $\models^{\mathsf{S}}_{\mathsf{IL}} \varphi$

The proof is given in the remainder of this subsection and will involve translating 'truth' in one semantics into 'truth' in the other semantics. The two directions of translation are covered in the next two lemmas.

This section is more technical, but the reasons for including it are the following two take-homes:

So, on a first read, focus only on the general ideas. Then proceed to more detail, e.g., explaining to yourself the formal definitions, etc.

1. Philosophically: This is another instance of the powerful idea of duality unifying the two different intuitions about semantics. Given a state-based model, construct an algebra of proposition given as truth-sets (cf. the first lemma below). Given an algebra, construct a state-based model by taking states to be appropriate models/filters (cf. the second lemma below).

2. Mathematically: The formal proofs and constructions constitute standard tools in logic that are worth learning.

Lemma 4.20 (From Kripke models to Heyting algebras)**.** *Let* $M = (S, R, I)$ *be an intuitionistic Kripke model. A subset* $U \subseteq S$ *is an* upset *if, for all* $s, s' \in S$ *with* $s \in U$ *and* sRs'*, also* $s' \in U$*.*

For this terminology, we think of R *as a pre-order. One can also say '*U *is closed under* R*'.*

Let $\mathsf{Up}(M)$ *be the set of upsets of* M*. For* $U, V \in \mathsf{Up}(M)$*, define*

$$U \to V := \big\{ s \in S : \textit{for all } s' \in S, \textit{if } sRs' \textit{ and } s' \in U, \textit{then } s' \in V \big\}.$$

As we'll see, this mirrors the clause for $\to$*.*

Define $v_M : \mathsf{P} \to \mathsf{Up}(M)$ *by* $v_M(p) = \{s \in S : s \vDash p\}$*. Then*

1. $(\mathsf{Up}(M), \cup, \cap, \to, \emptyset, S)$ *is a Heyting algebra and* v_M *a valuation.*

2. *For all* φ*,* $v_M(\varphi) = \{s \in S : s \vDash \varphi\}$*.*

Proof. Ad (1). The operations are well-defined: intersections and unions of upsets are again upsets, and both $\emptyset$ and S are upsets, so it remains to show that $U \to V$ is an upset: Indeed, if $s \in U \to V$ and sRs', show that $s' \in U \to V$. So let $s'' \in S$ with $s'Rs''$ and $s'' \in U$, and show $s'' \in V$. By transitivity, sRs'', so $s \in U \to V$ and $s'' \in U$ implies $s'' \in V$.

Since the set-theoretic operations $\cap$ and $\cup$ are distributive lattice operations with bounds $\emptyset$ and S, axiom 1 of Heyting algebras is satisfied. For axiom 2, first observe that $U \wedge (U \to V) \subseteq V$: if $s \in U$ and $s \in U \to V$, then, since sRs by reflexivity, $s \in V$. Next, if $W \in \mathsf{Up}(M)$ is such that $U \cap W \subseteq V$, then $W \subseteq U \to V$: Given $s \in W$, show $s \in U \to V$. So let $s' \in S$ with $s' \in U$ and sRs', and show $s' \in V$. Since $s \in W$ and sRs' and W is an upset, we have $s' \in W$. So $s' \in U \cap W \subseteq V$, as needed.

Finally, v_M is well-defined because, by the heredity condition, $\{s : s \vDash p\}$ is an upset and hence in $\mathsf{Up}(M)$.

Ad (2). By induction on φ. If φ is atomic, this holds by construction. For $\bot$, $\nu_M(\bot) = 0 = \emptyset = \{s : s \nvDash \bot\}$, similarly for $\top$. For $\psi \wedge \chi$,

$$\begin{aligned}\nu_M(\psi \wedge \chi) = \nu_M(\psi) \cap \nu_M(\chi) &\overset{\text{IH}}{=} \{s : s \vDash \psi\} \cap \{s : s \vDash \chi\} \\ &= \{s : s \vDash \psi \text{ and } s \vDash \chi\} = \{s : s \vDash \psi \wedge \chi\}.\end{aligned}$$

Similarly for $\vee$. For $\psi \to \chi$,

$$\begin{aligned}\nu_M(\psi \to \chi) = \nu_M(\psi) \to \nu_M(\chi) &\overset{\text{IH}}{=} \{s : s \vDash \psi\} \to \{s : s \vDash \chi\} \\ &= \{s : \forall s'.sRs' \;\&\; s' \vDash \psi \Rightarrow s' \vDash \chi\} = \{s : s \vDash \psi \to \chi\}.\end{aligned}$$

For $\neg$ and $\leftrightarrow$ use the definitions in terms of $\bot, \to, \wedge$. □

For the other direction, first recall that in a Heyting algebra A (or, more generally, a lattice), a subset $F \subseteq A$ is a *filter* if, for all $a, b \in A$, we have (a) if $a \in F$ and $a \leqslant b$, then $b \in F$ (upset), (b) if $a, b \in F$, then $a \wedge b \in F$ (closure), and (c) $F \neq \emptyset$ (nonempty). And F is *proper* if $0 \notin F$ (or, equivalently, $F \neq A$). The new idea is: A filter $F \subseteq A$ is *prime* if, for all $a, b \in A$,

If $a \vee b \in F$, then $a \in F$ or $b \in F$.

That's similar to ultrafilters (def. 2.24): Prime filters and ultrafilters are the same thing in Boolean algebras, but not so in Heyting algebras. There, prime filters are a more natural sense of 'model', since ultrafilters essentially require excluded middle.

Lemma 4.21 (From Heyting algebras to Kripke models)**.** *Let* $(A, \vee, \wedge, \to, 0, 1)$ *be a Heyting algebra that is nontrivial (i.e.,* A *isn't a singleton, or, equivalently,* $0 \neq 1$*). Let* $\nu : \mathsf{P} \to A$ *be a valuation. Let* S *be the set of proper prime filters on* A*. Let* $F \mathrel{R} F'$ *iff* $F \subseteq F'$*. Set* $I(F, p) = 1$ *iff* $\nu(p) \in F$*. Then*

1. $M(A) := (S, R, I)$ *is an intuitionistic Kripke model.*
2. *For all* φ*, we have* $M(A), F \vDash \varphi$ *iff* $\nu(\varphi) \in F$*.*

For the proof we use the following analogue of the Boolean Ultrafilter Theorem (exercise 2.c) for Heyting algebras:

(∗) If F is a filter of a Heyting algebra A and $a \in A \setminus F$, there is a prime filter F' such that $F \subseteq F'$ and $a \notin F'$.

Note F *is proper since* $a \notin F$

We omit the proof as it needs some tools that we don't cover here (namely Zorn's lemma; a statement equivalent to the Axiom of Choice in set theory).

For those interested, the sketch is: Consider the family $\mathcal{F}$ *of filters extending* F *without containing* a*. Every* $\subseteq$*-chain in* $\mathcal{F}$ *has an upper bound, namely the union of the filters in the chain. By Zorn's lemma,* $\mathcal{F}$ *has a maximal element* F' *which is the desired prime filter.*

Proof. Ad (1). First, S is an nonempty set: since A is non-trivial, it has a proper filter (e.g., $\uparrow 1 = \{1\} \not\ni 0$), so, by (∗), there is prime filter F' not containing 0, so $F' \in S$. Second, $\subseteq$ always is reflexive and transitive. Third, for the interpretation we have the heredity condition: if $F \subseteq F'$ and $I(F, p) = 1$, then $\nu(p) \in F$, so $\nu(p) \in F'$, so $I(F', p) = 1$.

Ad (2). By induction on φ. If φ is atomic, the claim holds by construction. For $\bot$, $F \vDash \bot$ never holds and $\nu(\bot) = 0 \notin F$ never holds (since F is proper). For $\top$, $F \vDash \top$ always holds and $\nu(\top) = 1 \in F$ always holds (since F is nonempty). For $\psi \wedge \chi$,

$$\begin{aligned} F \vDash \psi \wedge \chi \Leftrightarrow F \vDash \psi \text{ and } F \vDash \chi &\overset{\text{IH}}{\Leftrightarrow} \nu(\psi) \in F \text{ and } \nu(\chi) \in F \\ &\Leftrightarrow \nu(\psi) \wedge \nu(\chi) = \nu(\psi \wedge \chi) \in F, \end{aligned}$$

where the last step follows by 'closure' and 'upset' of F. For $\psi \vee \chi$, we reason similarly except now using 'upset' and 'prime' of F. It remains to show $F \vDash \psi \to \chi \Leftrightarrow \nu(\psi \to \chi) \in F$.

($\Leftarrow$) Assume $\nu(\psi \to \chi) \in F$ and show $F \vDash \psi \to \chi$. So let F' be a proper prime filter with $F \subseteq F'$ and $F' \vDash \psi$, and show $F' \vDash \chi$. By IH, $\nu(\psi) \in F'$. By assumption, $\nu(\psi) \to \nu(\chi) \in F \subseteq F'$. By 'closure' and axiom 2,

$$F' \ni \nu(\psi) \wedge \big(\nu(\psi) \to \nu(\chi)\big) \leqslant \nu(\chi).$$

By 'upset', $\nu(\chi) \in F'$.

($\Rightarrow$) Towards contradiction, assume $F \vDash \psi \to \chi$ and $\nu(\psi \to \chi) = \nu(\psi) \to \nu(\chi) \notin F$. Consider

$$G := \{a \in A : a \geqslant f \wedge \nu(\psi) \text{ for some } f \in F\}.$$

That's one of those 'consider' introducing a definition that comes out of the blue but magically works. Roll with it for now, and later come back to see why, in hindsight, it makes sense.

We check that G is a filter not containing $\nu(\chi)$: It is an upset by construction. If $a, a' \in G$ with $a \geqslant f \wedge \nu(\psi)$ and $a' \geqslant f' \wedge \nu(\psi)$, then $a \wedge a' \geqslant (f \wedge f') \wedge \nu(\psi)$ for $f \wedge f' \in F$, so $a \wedge a' \in G$. And G is nonempty since $1 \geqslant 1 \wedge \nu(\psi)$ with $1 \in F$. Finally, $\nu(\chi) \notin G$, since otherwise $\nu(\chi) \geqslant f \wedge \nu(\psi)$ for some $f \in F$, so, by the axiom on $\to$, $F \ni f \leqslant \nu(\psi) \to \nu(\chi)$, hence $\nu(\psi) \to \nu(\chi) \in F$.

Also note that $\nu(\psi) \in G$ (since $\nu(\psi) \geqslant 1 \wedge \nu(\psi)$ with $1 \in F$) and $F \subseteq G$ (if $f \in F$, then $f \geqslant f \wedge \nu(\psi)$, so $f \in G$).

So, by ($*$), we can extend G to a proper prime filter G' not containing $\nu(\chi)$. By IH and $\nu(\psi) \in G \subseteq G'$, we have $G' \vDash \psi$. Since $F \subseteq G \subseteq G'$ and $F \vDash \varphi \to \psi$, we have $G' \vDash \chi$. So IH implies $\nu(\chi) \in G'$, contradiction. □

Exercise 4.22. Use the two preceding lemmas to prove theorem 4.19. Hint: in the right-to-left direction you may find ($*$) helpful.

4.2.3 Soundness and completeness

That's a bonus subsection: it's not homework/exam material. (So we also allow ourselves to skip some details.) But it's there to illustrate the usefulness of algebraic semantics.

As already advertised, given the equivalence of the state-based and algebraic semantics, an advantage of the algebraic semantics is that it allows for a very elegant completeness proof. We'll do this here: provide a proof system for intuitionistic logic and show that a formula is derivable iff the formula is valid. (The direction 'derivable⇒valid' is know as soundness and the direction 'valid⇒derivable' is know as completeness.)

Definition 4.23. The (Hilbert type) *proof system* for IL is given as follows. The axioms are:

1. $\varphi \to (\psi \to \varphi)$
2. $\big(\varphi \to (\psi \to \chi)\big) \to \big((\varphi \to \psi) \to (\varphi \to \chi)\big)$
3. $\varphi \wedge \psi \to \varphi$
4. $\varphi \wedge \psi \to \psi$
5. $\varphi \to \big(\psi \to (\varphi \wedge \psi)\big)$
6. $\varphi \to \varphi \vee \psi$
7. $\psi \to \varphi \vee \psi$
8. $(\varphi \to \chi) \to \big((\psi \to \chi) \to (\varphi \vee \psi \to \chi)\big)$
9. $\bot \to \varphi$

A proof system for classical logic would be obtained by replacing axiom 9 with $\neg\neg\varphi \to \varphi$.

We write $\vdash_{\mathsf{IL}} \varphi$ if φ is an (substitution instance of) one of these axioms. The only inference rule is modus ponens: if one has derived φ (i.e., $\vdash_{\mathsf{IL}} \varphi$) and one has derived $\varphi \to \psi$ (i.e., $\vdash_{\mathsf{IL}} \varphi \to \psi$), then one can derive ψ (i.e., $\vdash_{\mathsf{IL}} \psi$). Here we treat $\neg, \top, \leftrightarrow$ as abbreviations for their defining formulas (e.g., $\neg\varphi := \varphi \to \bot$ and $\top := \neg\bot := \bot \to \bot$); otherwise we would need rules to state these equivalences.

Example 4.24. Here is a derivation of $\vdash_{\mathsf{IL}} \varphi \to \varphi$: Two instances of axiom 1:

$$\varphi \to \big((\varphi \to \varphi) \to \varphi\big)$$
$$\varphi \to (\varphi \to \varphi).$$

This illustrates that Hilbert systems are so minimal that no one actually uses them to prove things in the logic, but only to prove things about the proof system (like completeness).

An instance of axiom 2 is

$$\Big(\varphi \to \big((\varphi \to \varphi) \to \varphi\big)\Big) \to \Big(\big(\varphi \to (\varphi \to \varphi)\big) \to \big(\varphi \to \varphi\big)\Big).$$

So, after applying modus ponens twice, we get $\vdash_{\mathsf{IL}} \varphi \to \varphi$. ⌟

Given the proof system, we can construct a special Heyting algebra:

Definition 4.25. As before, let $\mathcal{L}_{\text{prop}}$ be the set of sentences of the propositional language. Two sentences $\varphi, \psi \in \mathcal{L}_{\text{prop}}$ are *provably equivalent*, written $\varphi \sim \psi$, if

$$\vdash_{\mathsf{IL}} \varphi \to \psi \text{ and } \vdash_{\mathsf{IL}} \psi \to \varphi.$$

It's not hard to show that $\sim$ is an equivalence relation, so we can identify the sentences that are provably equivalent. Formally, we consider the *quotient* $\mathcal{L}_{\text{prop}}/\sim = \{[\varphi]_\sim : \varphi \in \mathcal{L}_{\text{prop}}\}$ where $[\varphi]_\sim := \{\psi \in \mathcal{L}_{\text{prop}} : \varphi \sim \psi\}$ are the *equivalence classes*.

In fact, we can naturally define operations $\wedge, \vee, \to, 0, 1$ on $\mathcal{L}_{\text{prop}}/\sim$:

$$\begin{aligned} [\varphi]_\sim \wedge [\psi]_\sim &:= [\varphi \wedge \psi]_\sim & 0 &:= [\bot]_\sim \\ [\varphi]_\sim \vee [\psi]_\sim &:= [\varphi \vee \psi]_\sim & 1 &:= [\top]_\sim \\ [\varphi]_\sim \to [\psi]_\sim &:= [\varphi \to \psi]_\sim. \end{aligned}$$

One can again show that they are well-defined (e.g., if $\varphi \sim \varphi'$ and $\psi \sim \psi'$, then $\varphi \wedge \psi \sim \varphi' \wedge \psi'$). And one can show that they form a Heyting algebra

$$\mathsf{L} := (\mathcal{L}_{\text{prop}}/\sim, \vee, \wedge, \to, 0, 1)$$

which is called the *Lindenbaum–Tarksi algebra* (of intuitionistic logic).

This also works for many other logics: e.g., had we done it for classical logic, we would get a Boolean algebra.

And now we can prove soundness and completeness in a very elegant manner using the algebraic semantics. Given the equivalence of the semantics, this then also extends to the Kripke semantics.

Theorem 4.26. $\vdash_{\mathsf{IL}} \varphi$ *iff* $\vDash^{\mathsf{A}}_{\mathsf{IL}} \varphi$.

Proof. ($\Rightarrow$ aka soundness) By induction on derivations, show that if $\vdash_{\mathsf{IL}} \varphi$, then $\vDash^{\mathsf{A}}_{\mathsf{IL}} \varphi$: First, if φ is an instance of an axiom, then, it's readily checked that for any Heyting algebra A and valuation $\nu : \mathsf{P} \to \mathsf{A}$, one has $\nu(\varphi) = 1$. For example, for axiom 1, $\nu\big(\varphi \to (\psi \to \varphi)\big) = \nu(\varphi) \to \big(\nu(\psi) \to \nu(\varphi)\big) = 1$ since $1 \leqslant a \to (b \to a)$ iff $a = a \wedge 1 \leqslant b \to a$ iff $a \wedge b \leqslant a$ and the latter always holds. Second, one shows that this is preserved along modus ponens: if $\vDash^{\mathsf{A}}_{\mathsf{IL}} \varphi$ and $\vDash^{\mathsf{A}}_{\mathsf{IL}} \varphi \to \psi$, then $\vDash^{\mathsf{A}}_{\mathsf{IL}} \psi$. Indeed, if A is a Heyting algebra and $\nu : \mathsf{A} \to \mathsf{P}$ a valuation, we have, by assumption, $\nu(\varphi) = 1$ and $\nu(\varphi \to \psi) = 1$, so $1 = \nu(\varphi) \wedge 1 \leqslant \nu(\psi)$, so $\nu(\psi) = 1$.

($\Leftarrow$ aka completeness) Consider the Lindenbaum–Tarski algebra L and the valuation $\nu : \mathsf{P} \to \mathsf{L}$ defined by $\nu(p) := [p]_\sim$. It's readily seen, by

induction on φ, that $v(\varphi) = [\varphi]_\sim$. So $[\varphi]_\sim = v(\varphi) = 1 = [\top]_\sim$, so $\vdash_{\mathsf{IL}} \top \to \varphi$, so $\vdash_{\mathsf{IL}} \varphi$ (using modus ponens and $\vdash_{\mathsf{IL}} \top$). □

Two comments: First, notice that the proof actually shows that $\vdash_{\mathsf{IL}} \varphi$ iff in the Lindenbaum–Tarski algebra L with valuation v mapping p to $[p]$, we have $v(\varphi) = 1$. So intuitionistic logic also can, in a sense, be captured using a single algbera (of generalized truth-values), much like classical logic can be captured using only **2** (but, by Gödel's theorem, such an algebra needs to be infinite). Second, using the technique of 'canonical models' one can also show completeness for the Kripke semantics (e.g. Bezhanishvili and de Jongh 2006, sec. 3.3).

Intuitionistic vs. classical logic

Finally, let's compare intuitionistic logic to classical logic. On the one hand, any intuitionistic consequence $\Gamma \vDash_{\mathsf{IL}} \varphi$ also is a classical one $\Gamma \vDash_{\mathsf{CL}} \varphi$ (i.e., $\vDash_{\mathsf{IL}} \subseteq \vDash_{\mathsf{CL}}$). This can be seen using the algebraic semantics knowing that Boolean algebras are a special case of Heyting algebras. So in that sense classical logic is *stronger* than intuitionistic logic: any consequence of the latter is also achieved by the former. However, one might also point out that intuitionistic logic is stronger in another sense: it can see more differences between sentences than classical logic.

On the other hand, if φ is a classical validity, then $\neg\neg\varphi$ is an intuitionistic validity:

Theorem 4.27 (Glivenko)**.** $\vDash_{\mathsf{CL}} \varphi$ *iff* $\vDash_{\mathsf{IL}} \neg\neg\varphi$.

This fails for first-order intuitionistic logic, though!

We only sketch the proof: The right-to-left direction follows from $\vDash_{\mathsf{IL}} \subseteq \vDash_{\mathsf{CL}}$ and double negation elimination in classical logic. The other direction is either done by induction on the derivation of $\vDash_{\mathsf{CL}} \varphi$ or semantically by completeness of IL with respect to finite Kripke models and knowing that the final states of a Kripke model act like classical states.

4.2.4 Assessment

It is a delicate matter whether the intuitionistic motivation for logic is fully captured in the intuitionistic Kripke semantics or the algebraic semantics. There are many other semantics that also attempt to do this (e.g., formulas-as-types or Kleene's recursive realizabilities), though Brouwer was skeptical (for an overview, see Moschovakis 2021).

Nonetheless, the semantics surely get *some* things right: If, according to the BHK-interpretation, a proof of $\varphi \vee \psi$ is a proof of φ or a proof of ψ, we should have the so-called *disjunction property*: $\vdash \varphi \vee \psi$ iff $\vdash \varphi$ or $\vdash \psi$.

Example 4.28. We have: $\vDash_{\mathsf{IL}} \varphi \vee \psi$ iff $\vDash_{\mathsf{IL}} \varphi$ or $\vDash_{\mathsf{IL}} \psi$. Proof sketch: The right-to-left direction is clear. The other direction is by contraposition: Given Kripke models and states $M_0, s_0 \nVdash \varphi$ and $M_1, s_1 \nVdash \psi$, build a new Kripke model M by 'joining together' M_0 and M_1 and adding a new least state s_*. Then $M, s_* \nVdash \varphi \vee \psi$, because if $s_* \vDash \varphi$ (or $s_* \vDash \psi$), then, by heredity, also the higher-up state s_0 (resp., s_1) would make true φ (resp., ψ).

Note that this is blatantly false in classical logic: $\vDash_{\mathsf{CL}} p \vee \neg p$ but $\nVdash_{\mathsf{CL}} p$ and $\nVdash_{\mathsf{CL}} \neg p$. ⌟

Moreover, intuitionistic logic also inspired the so-called *Curry–Howard isomorphism* (aka *formulas-as-types paradigm*): that a formula (and its proofs) acts much like a type (and its constructable elements), as known in computer science. For example, a conditional $\varphi \to \psi$ acts like the type of functions from φ to ψ: a proof of $\varphi \to \psi$ is a function taking proofs of φ to proofs of ψ, so it describes an element of the function type $[\varphi \to \psi]$. For more on this, see the introductions of Sørensen and Urzyczyn (2006) and Troelstra (1992).

Following the 'meaning is use' idea, it also has been argued by Dummett and others that intuitionistic logic—rather than classical logic—is the quite generally correct logic. The idea is that the meaning of a sentence should not be described by when it corresponds to a fact in the real world (maybe fueled by skepticism about an external world as seen in section 4.2.1). Rather, it should be described by how it is used. So to know the meaning of a sentence is not to know when it is true but rather when it can be used, i.e., asserted—for which we need to be able to verify it. So the meaning of a sentence is linked to its possible 'proofs' which is line with the intuitionistic conception. For an overview and references, see Priest (2008, sec. 6.5 and 6.9).

4.3 Exercises

Exercise 4.a (Practice)**.** Do exercise 4.4 (2).

Exercise 4.b (Philosophical/problem)**.** In section 3.3.3, we discussed the issues of interpreting the 'definitely' operator Δ in the context of vagueness. If fuzzy logic is understood as a logic to reason with vague (or fuzzy)

statements, it arguably should interpret Δ. So it should provide a truth-function $\Delta : [0,1] \to [0,1]$ for the new unary connective Δ. Consider the function defined by $\Delta(x) := x^2$. In this exercise you should assess this suggested interpretation. To do so, you may consider the following and/or pursue your own ideas:

This is a common modeling for the hedge term 'very' in the fuzzy logic literature (Hájek 2001). For a more 'continuous' interpretation of the definite operator in the vagueness literature, see Pagin (2017).

- Why does this interpretation even make sense?
- Go through the discussion of section 3.3.3 and see how the above suggestion fares: which criticisms apply, which can it avoid?
- Consider some reasoning patterns that you think are philosophically crucial for a definite operator to hold or to fail. Say why you think they are crucial and prove/disprove them for the suggested interpretation. Some examples to consider:
  - $\vDash \Delta\varphi \to \varphi$
  - $\vDash \Delta(\varphi \to \psi) \to (\Delta\varphi \to \Delta\psi)$
  - $\vDash \Delta(\varphi \vee \psi) \to (\Delta\varphi \vee \Delta\psi)$
  - $\vDash \varphi \to \Delta\neg\Delta\neg\varphi$
  - The deduction theorem: $\Gamma, \varphi \vDash \psi$ iff $\Gamma \vDash \varphi \to \psi$
  - Contraposition: If $\neg\psi \vDash \neg\varphi$, then $\varphi \vDash \psi$.

For more philosophical exercises on fuzzy logic, see exercises 4.5–4.7 above.

Exercise 4.c (Problem)**.** In this exercise, we prove theorem 4.11. The key idea is to consider the following sentences

$$\varphi_n := \bigvee_{1 \leqslant k < l \leqslant n} \big((p_k \to p_l) \wedge (p_l \to p_k)\big).$$

Heyting's response continued: "In the sense of economy of thought this work is certainly useful, and in addition to that comes the particular beauty of your short proof." (see van Atten 2017, sec. 4.5.1).

1. Construct a Kripke model showing that, for any n, we have $\nvDash_{\mathsf{IL}} \varphi_n$.
2. Let L be a finitely-valued logic of the kind described in theorem 4.11: let T be the finite set of truth-values, $\mathsf{D} \subseteq \mathsf{T}$ the designated values, and $\neg, \wedge, \vee, \to, \leftrightarrow, \bot, \top$ the truth-functions. (So, e.g., for any L-valuation v, $v(\varphi \to \psi) = v(\varphi) \to v(\psi)$.) Let $\vDash_{\mathsf{L}}$ be its consequence relation, i.e., preservation of designated values under any L-valuation $v : \mathsf{P} \to \mathsf{T}$. And assume we have $\vDash_{\mathsf{IL}} = \vDash_{\mathsf{L}}$. Let $n := |\mathsf{T}| + 1$ (where $|\mathsf{T}|$ is the number of elements of T). Let $v : \mathsf{P} \to \mathsf{T}$ be an L-valuation. We show, in several steps, that $v(\varphi_n) \in \mathsf{D}$.

a) Show that there are $k < l$ such that $\nu(p_k) = \nu(p_l)$. Hint: use the pigeonhole principle.

b) Show that $\nu(p_k \to p_l) \in D$ and $\nu(p_l \to p_k) \in D$. Hint: show and use that $\vDash_{IL} p \to p$.

c) Show that $\nu\big((p_k \to p_l) \wedge (p_l \to p_k)\big) \in D$.

d) Show that $\nu(\varphi_n) \in D$.

3. Put together the previous two step to obtain a proof of theorem 4.11.

Exercise 4.d (Problem)**.** Show: If φ is a classical tautology with propositional atoms among $p_1, \dots, p_n$, then

$$\vDash_{IL} (p_1 \vee \neg p_1) \wedge \dots \wedge (p_n \vee \neg p_n) \to \varphi.$$

From Sørensen and Urzyczyn (2006, ex. 2.32 and 2.33).

Appreciate the intuitive statement: intuitionistic logic says that classical tautologies are okay once their atoms are decided.

Hint (and that's the hard part): Given a state s of a Kripke model M, say s *determines* p if either $M, s \vDash p$ or $M, s \vDash \neg p$; and define the classical valuation $\nu_s^M(q) = 1$ if $M, s \vDash q$ and $= 0$ otherwise. Show that if a state s of a model M determines all propositional atoms of φ, then $\nu_s^M(\varphi) = 1$ implies $M, s \vDash \varphi$.

Exercise 4.e (Problem)**.** Do exercise 4.22.

Exercise 4.f (Philosophical)**.** You can choose to write about one of the following topics:

1. Does the Kripke semantics do justice in formalizing the intuitionistic motivation?

2. Are you convinced by the 'meaning is use' argument for intuitionistic logic (cf. the last paragraph of section 4.2.4)?

In doing so, see if you can use the established duality (section Intuitionistic logic is infinitely-valued: duality) in building your philosophical arguments.

In general, again focus on a concrete aspect of the question that you find interesting, provide careful arguments, and consult the cited literature if you're looking for some inspiration.

4.4 Notes

Section 4.1 is based on Priest (2008, ch. 11). For a detailed book on the mathematics of fuzzy logic, see Hájek (1998). Section 4.2 is loosely based

on Bezhanishvili and de Jongh (2006), Priest (2008, ch. 6), Sørensen and Urzyczyn (2006, ch. 2), and the wikipedia article on Heyting algebras is also quite useful. For the historic development of intuitionistic logics, see van Atten (2017).

5 Hyperintensionality

Hyperintensionality became a popular concept in recent years. As a rule of thumb, it is used whenever one makes distinctions between things that are equivalent according to classical logic—or, more generally, are 'necessarily equivalent' (aka 'have the same intension'). A simple example is belief: If φ is a simple classical tautology and ψ is a very complicated one, we may believe φ but not ψ, even though the two sentences are necessarily equivalent. So a logic that is sensitive to belief needs to be more fine-grained than classical logic, thus yielding a non-classical logic.

In this chapter, we review the motivation for hyperintensional logics and semantics, discuss a particularly popular one (truthmaker semantics), and assess it by comparing it to other logics that we've seen.

Key concepts
- Extension vs. intension vs. hyperintension
- logical omniscience
- Exact vs. inexact truthmaking
- Exact truthmaker semantics
- Relation to previous logics (FDE, intuitionistic logic, weak Kleene)
- Subject matter
- Failure of distributivity, closure operators

5.1 Motivation

Possible worlds semantics is philosophy's success story which started in the second half of the 20th century. Many concepts that are central to philosophy—like meaning, belief, knowledge, or information—have been analyzed in terms of possible worlds. For example, the meaning of a sentence is analyzed as its *intension*, i.e., its truth-value profile across all possible worlds (formally, a function from possible worlds to classical truth-values). Or to believe a sentence is analyzed as the sentence being true in all possible worlds deemed possible ways the world could be. Here 'intensional' (not to be confused with 'inten*t*ional') is to be understood as opposed to 'extensional': rather than only looking at how things are, as a matter of fact, in this world (extensional), one also considers how things could have been otherwise (intensional).

The 'intensional revolution' of possible worlds semantics overcame the opposition from philosophers like Quine and Davidson who thought philosophy should be extensional (being doubtful about what these other possibilities really should be). Dissatisfied with the limited distinguishability of purely extensional approaches, it aimed to establish that there are meaningful intensional distinctions that can and should be made. Similarly, the 'hyperintensional revolution'—as it is sometimes put (Nolan 2014)—aims to establish that there are meaningful hyperintensional distinctions that can and should be made to overcome the too stringent purely intensional approach.

The reason of this dissatisfaction with intensional approaches is that, by using possible worlds, they cannot (at least not straightforwardly) distinguish necessarily equivalent sentences (i.e., sentences that are true at exactly the same possible worlds). And, in fact, many of the concepts analyzed by intensional approaches (like meaning, belief, etc.) are argued to really be hyperintensional. For example, if φ is a simple classical tautology and ψ a very complicated one, they are necessarily equivalent but we wouldn't want to say that believing the simple one entails believing the complicated one (*logical omniscience*). Here are some more examples.

1. There is a 40% chance of getting the job.

   There is a 60% chance of rejection.

Even though the two sentences are equivalent, they play a different cognitive role for us: upon hearing the first, we're more like to apply compared to the second (according to the work on framing effects by Kahneman and Tversky).

2. Mike is Mike.

   Mike is Jack the Ripper.

According to the received view on proper names in philosophy (since Kripke), the two sentences are necessarily equivalent. But epistemically they are very different: the first is no news, but knowing the latter may be very important.

3. The sun is shining.

   The sun is shining or both the sun is shining and it is raining.

The two sentences are logically equivalent according to classical logic. (In fact, they are equivalent according to any 'lattice-based' logics since this

basically is just the absorption law.) But they differ in 'topic' or 'subject matter': only one of them talks about rain.

But how fine-grained should we become? It seems like for any two syntactically distinct sentences φ and ψ, we can cook up a context in which they differ. (Maybe a rather non-smart AI that only stores the inputted sentences and hence doesn't 'believe' that φ and ψ actually mean the same thing.) So is the right amount of granularity simply syntactic identity?

Probably no: because in addition to the pressure to fine-grain from the above examples, there also is—maybe less prominently—pressure to coarse-grain (Bjerring and Schwarz 2017). For starters, there are clear synonyms: we would like the distinct sentences

4. They chill on the couch.

   They relax on the sofa.

to be identical in meaning. One reason might be that they communicate the same thing: if I utter one, I convey the same meaning if I utter the other. Another reason might be that they play the same cognitive role: believing or imagining one describes exactly the same states of mind as believing or imagining the other.

So the question is: Where on the 'continuum' between the 'coarse' extreme of intensional equivalence (or even extensional equivalence) and the 'fine' extreme of syntactic identity should the correct granularity be?

Of course, a truly philosophical answer first questions the presupposition of this question: can there even be a single correct granularity? Four comments:

1. It is reasonable to assume that the right granularity might be context-sensitive. For example, if we talk about a classical reasoner, $\neg\neg\varphi$ is synonymous to φ, but not so if we talk about an intuitionistic reasoner. In other words, the logic applicable within "…" in "Classically, …" is different from that of "Intuitionistically, …". Similarly, we might want to find the right logic of other sentential contexts (like "believe that …").

2. It might also be that there simply cannot be a correct granularity: if it gets one aspect right, it has to get another one wrong. (As argued for by Bjerring and Schwarz (2017).)

3. The previous chapters of these notes might also have primed us to think that 'correct granularity' is a vague concept: where it lies on

the continuum is not sharply defined.

4. Maybe the data against a purely intensional conception of meaning can be explained, so there is no need to go finer than that. For example, R. Stalnaker (1984) argues that the meaning of, say, a complicated tautology still simply is the set of all possible worlds, it is just that we don't know that this sentence has this meaning. (Though, we might still be interested in formalizing when we grasp that two sentences express the same proposition.) For more, see Berto and Nolan (2021, sec. 2.2).

Note the similarity to an epistemic theory of vagueness: there is a precise number of grains of sands needed to form a heap, it is just that we will never know which number it is.

Similarly to previous chapters, we won't defend here a particular answer to these questions. Rather, we look at logics that have been proposed that lie in between the two extremes. Actually any non-classical logic in the propositional language with a consequence relation finer than classical logic would qualify (e.g., strong Kleene, FDE, intuitionistic logic, etc.). But here we look at a particular prominent one for that purpose: truthmaker semantics.

5.2 Formal logic: truthmaker semantics

There are many formal approaches to hyperintensionality: see Berto and Nolan (2021, sec. 4). Some are more fine-grained than others. Here we focus on a particularly popular one: truthmaker semantics (Fine 2017). This again comes in various forms, here we look at the finest one, namely, exact truthmaker semantics (Fine and Jago 2019).

Okay, sorry, pun intended.

5.2.1 Exact truthmaker semantics

The idea of a truthmaker is prominent in contemporary philosophy—going back to the 1980s (MacBride 2021). When saying "x makes true p", one wants to express that some worldly thing x (e.g., a fact, situation, or state of affairs) is the reason why the linguistic thing p (e.g., a sentence or proposition) is true. And one speaks of the *truthmaker* x and the *truthbearer* p.

One can distinguish two 'dual' perspectives on truthmakers (Fine 2017): The metaphysic perspective focuses on *what* truthmakers are, regardless of how they make true sentences, hoping to learn something about the world from our knowledge of language. The semantic perspective focuses on *how* truthmakers make true sentences, regardless of what they are, hoping

to learn something about language from our knowledge about the world. Here we pursue the semantic perspective.

So, how should truthmaking work? Consider, as a start, a possible world (or classical state) s in which it rains. So s makes true the sentence $p =$ 'it is raining'. But since s is complete, it is also decided on any other sentence, so s contains a lot of facts that are not relevant to making true the sentence p. So, as far as truthmaking is concerned, we can also consider incomplete (and possibly also inconsistent) states. For example, the state s described by it raining hence still is a truthmaker for p.

In fact, state s is an *exact* truthmaker for p: it contains just as much as is needed to make true p. The state s' described by it raining and being windy is only an *inexact* truthmaker for p: it still makes true p but it contains more than needed for this (namely that it also is windy). In other words, while state s is *wholly* relevant to p, state s' is only *partially* relevant to p.

However, the state s' of it raining and being windy is an exact truthmaker for the conjunction $p \wedge q$ where $q =$ 'it is windy'. This is because s' is precisely the fusion of its two parts s and s'' with s being the state of it raining which exactly makes true p and s'' being the state of it being windy which exactly makes true q.

These ideas are now captured in the following (formal) exact truthmaker semantics. We follow template 2.8 for state-based semantics. For simplicity, we work in this chapter only with the language built from propositional atoms using only $\neg, \wedge, \vee$.

Definition 5.1. First, a *truthmaker model* M is a triple $(S, \leqslant, I)$ such that

- S is a set (whose elements are called states).
- $\leqslant$ is a partial order such that any two $s, s' \in S$ have a least upper bound denoted $s \sqcup s'$ (also called the fusion of s and s').

Sometimes it is assumed that $(S, \leqslant)$ is complete, i.e., every subset of S has a least upper bound (not just pairs of states).

- $I = (I^+, I^-)$ is a pair of functions $S \times \mathsf{P} \to \{0, 1\}$ satisfying $\sqcup$-*closure*:

  If $I^+(s, p) = 1$ and $I^+(s', p) = 1$, then $I^+(s \sqcup s', p) = 1$

  If $I^-(s, p) = 1$ and $I^-(s', p) = 1$, then $I^-(s \sqcup s', p) = 1$.

  If $I^+(s, p) = 1$ (resp., $= 0$), we say s makes true p (resp., doesn't make true p), and if $I^-(s, p) = 1$ (resp., $= 0$), we say s makes false p (resp., doesn't make false p).

Second, we extend truthmaking ($\vDash^+$) and falsemaking ($\vDash^-$) to all formulas:

This version of the semantics is sometimes called 'inclusive', while the 'non-inclusive' version doesn't add $s \vDash^- \varphi \vee \psi$ to the clause for $\vDash^- \varphi \wedge \psi$ (similarly for $\vee$).

- $s \vDash^+ p$ iff $I^+(s,p) = 1$

  $s \vDash^- p$ iff $I^-(s,p) = 1$.

- $s \vDash^+ \neg\varphi$ iff $s \vDash^- \varphi$

  $s \vDash^- \neg\varphi$ iff $s \vDash^+ \varphi$

- $s \vDash^+ \varphi \wedge \psi$ iff there are $s', s'' \in S$ with $s' \sqcup s'' = s$ and $s' \vDash^+ \varphi$ and $s'' \vDash^+ \psi$.

  $s \vDash^- \varphi \wedge \psi$ iff $s \vDash^- \varphi$ or $s \vDash^- \psi$ or there are $s', s'' \in S$ with $s' \sqcup s'' = s$ and $s' \vDash^- \varphi$ and $s'' \vDash^- \psi$.

- $s \vDash^+ \varphi \vee \psi$ iff $s \vDash^+ \varphi$ or $s \vDash^+ \psi$ or there are $s', s'' \in S$ with $s' \sqcup s'' = s$ and $s' \vDash^+ \varphi$ and $s'' \vDash^+ \psi$.

  $s \vDash^- \varphi \vee \psi$ iff there are $s', s'' \in S$ with $s' \sqcup s'' = s$ and $s' \vDash^- \varphi$ and $s'' \vDash^- \psi$.

Third, we define *exact truthmaking consequence* $\Gamma \vDash_{\mathsf{TM}} \varphi$ as truthmaking-preservation: For all truthmaker models M and states s, if $s \vDash^+ \psi$ for all $\psi \in \Gamma$, then $s \vDash^+ \varphi$. Two sentences φ and ψ are equivalent, written $\varphi \dashv\vDash \psi$, if $\varphi \vDash_{\mathsf{TM}} \psi$ and $\psi \vDash_{\mathsf{TM}} \varphi$.

It's instructive to compare this semantics to the truthmaking and falsemaking semantics for FDE from exercise 3.d. They are similar in that both allow states that may be incomplete or inconsistent. But they differ in that the FDE semantics may be said to be *extensional*: to determine whether a state makes true or false a formula you only have to consider that single state and no other states. While the truthmaker semantics is *intensional* or *modal*: to determine whether a state makes true, for example, a conjunction, you also need to consider other states, namely those states that, when fused, yield the current state. This is why we now need an additional relation on the states space (the order $\leqslant$ and the resulting fusion $\sqcup$), which wasn't needed for the FDE semantics.

In the Kripke semantics for modal logic, conjunction is extensional, but the connective $\Box$ is modal: to determine whether a state makes true $\Box\varphi$ you also need to consider other states.

The FDE semantics has the spirit of inexact truthmaking, as can be seen in the clause for conjunction. For exact truthmaking, the clause "$s \vDash^+ \varphi\wedge\psi$ iff $s \vDash^+ \varphi$ and $s \vDash^+ \psi$" doesn't make sense: if s is an exact truthmaker for φ, it contains just as much to make true φ, so it cannot contain additional or different stuff to make true ψ (given ψ 'talks about' other things than φ), so s cannot be an exact truthmaker of ψ. However, for inexact truthmaker this clause is the natural choice (hence also used by the FDE semantics): if a state s inexactly makes true φ and inexactly makes true ψ, it contains

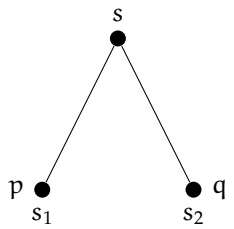

Figure 5.1: A truthmaker model with $s \vDash^+ p \wedge q$ without $s \vDash^+ p$ or $s \vDash^+ q$.

something that makes true φ and something that makes true ψ, so it inexactly makes true $\varphi \wedge \psi$.

This difference is illustrated in the following example.

Example 5.2. Consider the truthmaker model described in figure 5.1 (convince yourself that it indeed describes one). There we have $s_1 \vDash^+ p$ and $s_2 \vDash^+ q$, so, since $s = s_1 \sqcup s_2$, also $s \vDash^+ p \wedge q$. However, neither $s \vDash^+ p$ nor $s \vDash^+ q$. That's a rather weird feature for a logic, but a consequence of the exactness interpretation.

In particular, unlike many other logics, the 'distributive entailment' $\varphi, \psi \vDash_{TM} \varphi$ is *not* the same as the 'conjunctive entailment' $\varphi \wedge \psi \nvDash_{TM} \varphi$. ⌟

The following exercise establishes some equivalences in the logic.

Exercise 5.3.

1. De Morgan: Show $\neg(\varphi \wedge \psi) \dashv\vDash \neg\varphi \vee \neg\psi$. And $\neg(\varphi \vee \psi) \dashv\vDash \neg\varphi \wedge \neg\psi$.

2. Not distributive: Show $\varphi \wedge (\psi \vee \chi) \dashv\vDash (\varphi \wedge \psi) \vee (\varphi \wedge \chi)$. Show $\varphi \vee (\psi \wedge \chi) \vDash_{TM} (\varphi \vee \psi) \wedge (\varphi \vee \chi)$. Provide a simple countermodel to the other direction.

3. Compare this situation to that of intuitionistic logic (hint: it's precisely the other way round). (Though there is a version of truthmaker semantics for intuitionistic logic, see below.)

4. $\varphi \wedge (\varphi \vee \psi) \dashv\vDash \varphi \vee (\varphi \wedge \psi)$, but neither sentence is equivalent to just φ.

Finally, we note that the closure condition extends to all formulas. This is analogous to intuitionistic logic: there we've demanded the heredity condition (being closed under R) for all atomic formulas and showed that it extends to all formulas. For truthmaker semantics we did *not* demand closure under the relation $\leqslant$ (and it's, e.g., violated in figure 5.1), but under the fusion $\sqcup$.

Proposition 5.4 (Closure). *For all formulas* φ*, all truthmaker models* M*, and states* s *and* s'*:*

- *If* $s \vDash^+ \varphi$ *and* $s' \vDash^+ \varphi$*, then* $s \sqcup s' \vDash^+ \varphi$*.*
- *If* $s \vDash^- \varphi$ *and* $s' \vDash^- \varphi$*, then* $s \sqcup s' \vDash^- \varphi$*.*

Proof. By induction on φ. For $\varphi = p$ atomic, this is the closure assumption.

For $\neg\varphi$, if $s \vDash^+ \neg\varphi$ and $s' \vDash^+ \neg\varphi$, then $s \vDash^- \varphi$ and $s' \vDash^- \varphi$, so, by IH, $s \sqcup s' \vDash^- \varphi$, so $s \sqcup s' \vDash^+ \neg\varphi$. Similarly for $\vDash^-$.

For $\varphi \wedge \psi$, first consider $\vDash^+$. If $s \vDash^+ \varphi \wedge \psi$ and $s' \vDash^+ \varphi \wedge \psi$, then there are t_1, t_2, u_1, u_2 with

$$\begin{array}{lll} s = t_1 \sqcup t_2 & t_1 \vDash^+ \varphi & t_2 \vDash^+ \psi \\ s' = u_1 \sqcup u_2 & u_1 \vDash^+ \varphi & u_2 \vDash^+ \psi. \end{array}$$

By IH, $t_1 \sqcup u_1 \vDash^+ \varphi$ and $t_2 \sqcup u_2 \vDash^+ \psi$. So $s = (t_1 \sqcup u_1) \sqcup (t_2 \sqcup u_2) \vDash^+ \varphi \wedge \psi$, as needed.

Second, consider $\vDash^-$. If $s \vDash^- \varphi \wedge \psi$ and $s' \vDash^- \varphi \wedge \psi$, then either (i) $s \vDash^- \varphi$, or (ii) $s \vDash^- \psi$, or (iii) there is t_1, t_2 with $s = t_1 \sqcup t_2$ and $t_1 \vDash^- \varphi$ and $t_2 \vDash^- \psi$ and either (i)' $s' \vDash^- \varphi$, or (ii)' $s' \vDash^- \psi$, or (iii)' there is u_1, u_2 with $s' = u_1 \sqcup u_2$ and $u_1 \vDash^- \varphi$ and $u_2 \vDash^- \psi$. Let's consider the possible combinations of primed and unprimed cases. If (i) and (i)', then, by IH, $s \sqcup s' \vDash^- \varphi$, so $s \sqcup s' \vDash^- \varphi \wedge \psi$. If (i) and (ii)', then, by definition, $s \sqcup s' \vDash^- \varphi \wedge \psi$. If (i) and (iii)', then, by IH, $s \sqcup u_1 \vDash^- \varphi$ and, so, by definition, $s \sqcup s' = (s \sqcup u_1) \sqcup u_2 \vDash^- \varphi \wedge \psi$. The cases (ii)&(i)', (ii)&(ii)', and (ii)&(iii)' are analogous. As are (iii)&(i)' and (iii)&(ii)'. If (iii) and (iii)', then, by IH, $t_1 \sqcup u_1 \vDash^- \varphi$ and $t_2 \sqcup u_2 \vDash^- \psi$, so $s \sqcup s' = (t_1 \sqcup u_1) \sqcup (t_2 \sqcup u_2) \vDash^- \varphi \wedge \psi$.

For $\varphi \vee \psi$ we use the De Morgan laws (exercise 5.3 (2)). □

5.3 Assessment

Coming back to the question of what the right granularity of meaning should be, we have now seen one logic between the two extremes of intensional equivalence and syntax. As mentioned, the logics that we've seen previously also lie within this continuum. So what's the difference?

One big difference concerns *subject matter*. Many of the previous logics are 'lattice-based': conjunction and disjunction are interpreted as the operations $\wedge$ and $\vee$ of a lattice. As a result of the absorption law, $p \wedge (p \vee q)$ is equivalent to p (and similarly $p \vee (p \wedge q)$ is, too, equivalent to p). However, these sentences differ in subject matter: the (topic of) sentence q only

occurs in the first sentence (also see example 3 of section 5.1). On the other hand, in exercise 5.3 (4), we've seen that truthmaker semantics can distinguish $p \wedge (p \vee q)$ from p (and also $p \vee (p \wedge q)$ from p). In particular, truthmaker semantics is not lattice-based.

Exercise 5.5 (Comparison with weak Kleene)**.**

1. The only non-lattice-based logic that we've seen was weak Kleene K_3^w: and it, too, doesn't make equivalent p and $p \wedge (p \vee q)$ (and also not p and $p \vee (p \wedge q)$): Intuitively, if the topic of conversation includes p but not q, then if p is true, it has value 1 because it is true-and-on-topic, while both $p \wedge (p \vee q)$ and $p \vee (p \wedge q)$ are $\mathfrak{i}$ because they're off-topic (since they talk about something that's not the topic of discussion). Turn this into a formal argument.

2. In fact, show that φ and ψ have the same value under any weak Kleene valuation iff φ and ψ are classically equivalent and $\mathsf{At}(\varphi) = \mathsf{At}(\psi)$.

3. Can you find two formulas that are equivalent in K_3^w but not in truthmaker semantics?

In fact, truthmaker semantics can provide a positive account—i.e., a formalization—of subject matter. (As opposed to just negatively saying when two sentences are distinct in subject matter because of having different atoms.) The subject matter $\sigma(\varphi)$ of a sentence φ (with respect to a truthmaker model) is the fusion $s_1 \sqcup s_2 \sqcup \dots$ of the truthmakers $s_1, s_2, \dots$ of φ. (Technically, this requires the order $(S, \leqslant)$ to be complete.) Thus, subject matters are states (which is an identification worth discussing). But this has some neat consequences that one would expect: the subject matter of a conjunction $\varphi \wedge \psi$ is the same as the subject matter of the corresponding disjunction $\varphi \vee \psi$. That's known as *transparency* (or *junction*) and reflects the idea that logic is topic neutral: logical connectives don't add any subject matter to a sentence. Formally:

This makes sense: What is φ about? Well, it's about (the fusion of) all the situations in which it is true! (Think about whether this would this work for possible worlds, too.)

Exercise 5.6. Let $M = (S, \leqslant, I)$ be a truthmaker model. For a formula φ, define the truth-set $|\varphi|^+ = \{s \in S : s \vDash^+ \varphi\}$ and falsity-set $|\varphi|^- = \{s \in S : s \vDash^- \varphi\}$. Assume

- $(S, \leqslant)$ is complete, i.e., every subset $A \subseteq S$ has a least upper bound, denoted $\bigvee A$.
- For all atoms p, $|p|^+$ and $|p|^-$ are nonempty. (This entails that the truth- and falsity-set of any formula φ are nonempty.)

This is a common assumption (e.g. Fine 2016). Define the subject matter $\sigma(\varphi)$ of a formula φ as

$$\sigma(\varphi) := \bigvee |\varphi|^{+}.$$

Show that $\sigma(\varphi \wedge \psi) = \sigma(\varphi) \sqcup \sigma(\psi) = \sigma(\varphi \vee \psi)$.

Also, since subject matters are states, the mereology (i.e., notion of parthood) for states given by $\leqslant$ and $\sqcup$ also provides a mereology for subject matter. Further, one can restrict a proposition $A \subseteq S$ to a subject matter s obtaining the proposition $\{a \sqcap s : a \in A\}$ where $s \sqcap s'$ is the fusion of all states that are a common parts of s and s'. See Fine (2017, Part II) for more applications of truthmaker semantics to philosophy and linguistics (partial content, counterfactuals, imperatives, and scalar implicatures).

The general fact behind this is that in a complete partial order also every subset has a greatest lower bound: namely the greatest upper bound of all its lower bounds.

Another comparison to logics that we've seen so far is with intuitionistic logic. Fine (2014) shows how a version of truthmaker semantics can provide a semantics for intuitionistic logic that is, in a way, a common semantic framework behind both classical and intuitionistic logic.

Finally, one might wonder: how 'hyperintensional' is truthmaker semantics? It is hyperintensional in the (standard) sense that it can distinguish between necessarily equivalent sentences. However, it is *not* hyperintensional in the stronger sense that its distinguishability is beyond any intensional approach: van Benthem (2017) shows that truthmaker semantics can also be viewed as bimodal logic (which hence operates on possible worlds only).

5.4 Exercises

In the following exercises, we follow the general explanation of Restall (2000, ch. 12) for failures of distributivity and apply it to truthmaker semantics. Afterward, we end with a proof of compactness and a philosophical question.

There's a chance we're doing something new here: I haven't seen that application. (But also didn't check thoroughly; though cf. Fine (2016, lem. 6).)

Exercise 5.a (Practice)**.** Assume a state-based semantics uses, among others, the following definitions for models M and states s:

1. $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$
2. $M, s \vDash \varphi \vee \psi$ iff $M, s \vDash \varphi$ or $M, s \vDash \psi$
3. $\varphi \vDash \psi$ iff for all models M and states s, if $M, s \vDash \varphi$, then $M, s \vDash \psi$.

Show that then distributivity already holds: $\varphi \vee (\psi \wedge \chi) \dashv\vDash (\varphi \vee \psi) \wedge (\varphi \vee \chi)$.

Note that the semantics for FDE *in exercise 3.d has these features and* FDE *indeed satisfies distributivity.*

So if distributivity should fail, at least one of the features needs to go. In truthmaker semantics, the first two go:

Exercise 5.b (Practice)**.** Show that truthmaker semantics violates both (1) and (2), and that it violates distributivity (cf. exercise 5.3 (2)).

For other non-distributive logics (e.g., quantum logics), a popular choice is to give up (2). Here are two motivations for this.

First, algebraic: From the algebraic perspective, 'models' of the logic (which are much like states in a corresponding state-based semantics) are certain filters of the algebras corresponding to the logic. For classical logic, these were ultrafilters in Boolean algebras. For intuitionistic logic, these were prime filters in Heyting algebras. For both types of filters F, we have $a \vee b \in F$ iff $a \in F$ or $b \in F$. A natural generalization would be to consider any filter (not necessarily ultra or prime). But then, in general, we don't have $a \vee b \in F \Rightarrow a \in F$ or $b \in F$ anymore. Thinking of F as a state, it would violate (2).

Second, state-based: In a state-based semantics, we often want that the 'truth-set' $[\![\varphi]\!] = \{s : s \vDash \varphi\}$ of formulas have certain closure properties. In intuitionistic logic, they should be upsets (closed under the relation R). In truthmaker semantics, they should be closed under fusion $\sqcup$. In intuitionistic logic, we were lucky that unions of upsets are again upsets, so we could define $s \vDash \varphi \vee \psi$ iff $s \vDash \varphi$ or $s \vDash \psi$, hence $[\![\varphi \vee \psi]\!] = [\![\varphi]\!] \cup [\![\psi]\!]$ is again an upset. However, in truthmaker semantics, this doesn't work: the union of two set of states closed under fusion need not be closed under fusion (take, e.g., two singletons $\{s\}$ and $\{s'\}$ with s and s' $\leqslant$-incomparable). So we should take $[\![\varphi \vee \psi]\!]$ to be the 'smallest' set closed under fusion which contains $[\![\varphi]\!] \cup [\![\psi]\!]$. But then, if the closure of $[\![\varphi]\!] \cup [\![\psi]\!]$ wasn't trivial, it contains a new state s that wasn't already in $[\![\varphi]\!] \cup [\![\psi]\!]$. And for this state we have $s \vDash \varphi \vee \psi$ (since s is in the closure) but $s \nvDash \varphi$ and $s \nvDash \psi$ (since s is not in the union), thus violating (2).

In quantum logic, the truth-sets should be closed subspaces of an underlying Hilbert space describing the states of the quantum system. But unions of closed subspaces aren't necessarily closed subspaces again, so one needs to close the union (Dalla Chiara 1986).

So, to avoid (2)—maybe as part of a general attempt to get a logic without distributivity—, one would define: $s \vDash \varphi \vee \psi$ iff s is in the closure of $[\![\varphi]\!] \cup [\![\psi]\!]$. We now see how to do this precisely for truthmaker semantics.

Exercise 5.c (Problem)**.** Let $M = (S, \leqslant, I)$ be a truthmaker model. (The interpretation I is not important in this exercise, so we could just consider the underlying frame $(S, \leqslant)$.) For a subset $A \subseteq S$ define

$$\mathcal{C}(A) := \{a_1 \sqcup \ldots \sqcup a_n : a_1, \ldots, a_n \in A, n \geqslant 1\}.$$

So $\mathcal{C}$ is a function $\mathcal{P}(S) \to \mathcal{P}(S)$ (where $\mathcal{P}(S)$ is the powerset of S, i.e., the set of all subsets of S). Show

1. $\mathcal{C}(A)$ is the least subset of S which contains A and is closed under $\sqcup$: i.e.,
   a) $A \subseteq \mathcal{C}(A)$ and if $s, s' \in \mathcal{C}(A)$, then $s \sqcup s' \in \mathcal{C}(A)$, and
   b) if $B \subseteq S$ is such that $A \subseteq B$ and $s, s' \in B$ implies $s \sqcup s' \in B$, then $\mathcal{C}(A) \subseteq B$.

Conclude that $\mathcal{C}$ is a *closure operator* on S, i.e., it has the following properties:

2. Increasing: $A \subseteq \mathcal{C}(A)$
3. Idempotent: $\mathcal{C}\mathcal{C}(A) = \mathcal{C}(A)$
4. Monotone: If $A \subseteq B$, then $\mathcal{C}(A) \subseteq \mathcal{C}(B)$.

(If $\mathcal{C}(A) = A$, one says that A is *closed*. Closure operators are, albeit quite abstract, an important concept in logic, algebra, and topology: if you want to know more, a good start is the wikipedia article.)

The closure of a set in topology is an example of a closure operator. But it satisfies more properties: the union of closed sets is closed, while this need not be true for general closure operators.

Exercise 5.d (Problem)**.** Let $M = (S, \leqslant, I)$ be a truthmaker model. Define $[\![\varphi]\!]^+ := \{s \in S : s \vDash^+ \varphi\}$. Show $s \vDash^+ \varphi \vee \psi$ iff $s \in \mathcal{C}\big([\![\varphi]\!]^+ \cup [\![\psi]\!]^+\big)$.

We end with two further exercises: a difficult one on compactness and a philosophical one.

Exercise 5.e (Problem)**.** This exercise establishes the compactness theorem for truthmaker semantics: For any (possibly infinite) set of formulas Γ and formula φ, if $\Gamma \vDash_{\mathsf{TM}} \varphi$, then there is a finite subset Γ_0 of Γ such that $\Gamma_0 \vDash_{\mathsf{TM}} \varphi$. (The other direction is trivial.)

Hint: By contraposition, writing X for the finite subsets of Γ, assume there is, for every $i \in X$, a truthmaker model $M_i = (S_i, \leqslant_i, I_i)$ and state $s_i \in S_i$ with $s_i \vDash^+ \psi$ for every $\psi \in i$ but $s_i \nvDash^+ \varphi$. To merge the M_i and s_i into a truthmaker model $M = (S, \leqslant, I)$ and state s witnessing $\Gamma \nvDash_{\mathsf{TM}} \varphi$, first set

$$\mathcal{F} := \big\{A \subseteq X : \exists i \in X\, \forall j \in X (i \subseteq j \Rightarrow j \in A)\big\}.$$

This proof sketch is based on the ultraproduct proof (my favorite one) for the compactness theorem of first-order logic. So, more ultrafilter magic :-) It's more semantic than the more syntactic proof of Fine and Jago (2019). (Again, I haven't seen the proof before.)

Show this is a proper filter on the Boolean algebra $\mathcal{P}(X)$. You may refer to (a version of) the Boolean Ultrafilter Theorem to extend $\mathcal{F}$ to an ultrafilter $\mathcal{U}$. Now take $(S, \leqslant)$ to be the product of the $(S_i, \leqslant_i)$ over X and $I^+(s, p) = 1$ iff $\{i \in X : M_i, s(i) \vDash^+ p\} \in \mathcal{U}$ (similarly for $\vDash^-$). Prove it has the required properties by showing that the definition of $s \vDash^{\pm} p$ extends to all formulas (you may use the De Morgan laws from exercise 5.3 1).

Exercise 5.f (Philosophical)**.** You can choose to write about one of the following topics:

1. In exercises 5.a–5.d, you've motivated and showed that truthmaker semantics gives up on the standard clause for disjunction (2) and instead adds a closure operator. Reflect on the philosophical significance of this. Moreover, as you've shown, truthmaker semantics also gives up on the standard clause for conjunction (1). What to make of this philosophically? (Also cf. the inclusive vs. non-inclusive version of truthmaker semantics.)

2. Should we stick with possible worlds semantics and explain away the data for hyperintensional distinctions differently?

3. Think about the two sentences p and $p \vee (p \wedge q)$ (or $p \wedge (p \vee q)$). Truthmaker semantics is fine-grained enough to make them non-equivalent following the idea that synonymy should entail subject matter identity. Do you agree or do you have reason to coarse-grain (like lattice-based logics making them equivalent)?

4. Explore the idea from section 5.1 of "the correct granularity" being a vague predicate. For example, on a negative note, does this run counter the idea that the background logic under which we operate is fixed? Or, on a positive note, can this account for sorites-like sequences of sentences $\varphi_1, \ldots, \varphi_n$ where adjacent sentences are considered synonymous while the change in meaning from φ_1 to φ_n is too much for them to be considered synonymous? (Since 'meaning identity' is transitive, this is hard to account for otherwise.)

Again, focus on a concrete aspect of the question that you find interesting, provide clear and careful arguments, and consult the cited literature if you look for some inspiration.

5.5 Notes

Loosely following Berto and Nolan (2021) for the more philosophical parts and Fine and Jago (2019) for the exact truthmaker semantics (and also Fine (2014, 2016, 2017)). (Though, for full disclosure, I can't deny Hornischer (2017, 2020).)

6 Counterfactuals

In this and the next chapters, we focus on conditionals: sentences of the form 'if φ, then ψ'. They are omnipresent in natural language and everyday reasoning, so it would be good to have a theory about them—though that's notoriously difficult. In this chapter, we look at one big class of conditionals: counterfactuals, i.e., conditionals of the form 'if φ were the case, then ψ would be the case'. In the next chapter, we look at other classes.

Although we've already seen several formal conditionals in previous logics, we'll see that none of them can model counterfactuals. Instead, we introduce a state-based semantics capturing the idea that a counterfactual is true if the closest states making the antecedent true also make the consequent true. We then both logically and philosophically analyze this semantics.

Key concepts
- Indicative vs. subjunctive/counterfactual conditionals
- Strengthening the antecedent
- Intuitive idea for counterfactuals: closest antecedent-worlds are consequent-worlds
- Formal semantics for counterfactuals using similarity models
- Limit assumption, how it simplifies the semantics, and violates compactness
- Correspondence theory: modus ponens vs. weak centering, rational monotonicity vs. almost connected
- Criticisms of the similarity relation

6.1 Motivation

We use conditionals all the time in communication and reasoning. So it would be good to have a general understanding of their logic. The problem is, however, that conditionals are very puzzling and come in different kinds, which makes it difficult to build a unified theory. Let's collect some of those difficulties.

First, one usually distinghuishes two types of conditionals in natural language: *Indicative* conditionals like

1. If Oswald didn't shoot Kennedy, then someone else did.

And *subjunctive* or *counterfactual* conditionals (those with a 'would') like

2. If Oswald hadn't shot Kennedy, then someone else would have.

Despite both having the 'if, then' shape, they cannot both get the same formal semantics, since (1) is true while (2) is false. In this chapter, we look at counterfactuals and later at other, i.e., indicative, conditionals. Moreover, it's not just their omnipresence that makes counterfactuals a worthwhile object of study. They also are a useful tool in philosophy to formulate theories: for example, analyses of causation (Lewis) or knowledge (Nozick). So it's also worth better understanding counterfactuals from this 'metaphilosophical' point of view.

Second, generally speaking, counterfactuals are of the form

> If it had been the case that φ, then it would have been the case that ψ.

So they reason, as the name sugests, about situations different from the actual situation. In particular, we cannot model them with the material conditional $\neg\varphi \vee \psi$: since the antecedent of a counterfactual is (typically) false, the material conditional would be true regardless of what the consequent says (so this would make (2) true). (There are more so-called *paradoxes of the material conditional* which are taken as arguments that it also is not a good model for indicative conditionals, this will be the topic of chapter 8.)

Third, consider the following inference pattern which is satisfied by many formal conditionals:

$$\varphi \to \psi \vDash \varphi \wedge \chi \to \psi \qquad \text{(Strengthening the Antecedent)}$$

However, this pattern is violated by counterfactuals: arguably, we have

3. If I had taken the bike (instead of walking), I would have been faster.

But we don't have

4. If I had taken the bike and had a puncture, I would have been faster.

So this rules out any formal conditional satisfying strengthening of the antecedent.

Exercise 6.1. Go through some logics with a conditional that you already know and check whether they satisfy strengthening of the antecedent.

Hint: this includes classical logic (i.e., the material conditional), intuitionistic logic, fuzzy logic, and strong Kleene. It is violated in weak Kleene.

So most formal tools that we know so far are a non-starter for counterfactuals. What, then, is a promising approach? The insight of Lewis, Stalnaker, and others was the idea that:

See Starr (2019) for a history of this idea.

(∗) a counterfactual $\varphi \to \psi$ should be true at a world s iff in the worlds s' that make φ true and are most similar to s, also ψ is true.

Even though this is an informal idea, it's intuitive enough to see that (3) indeed does not imply (4): Consider the world s' (resp., s'') which is like the actual world s but where I cycled (resp., and also got a puncture) instead of walked. Now s' is the closest world where the antecedent of (3) is true. There is no reason to assume that anything is wrong with the bike: otherwise the world would be more dissimilar to the actual world where the bike is fine. So, in s', I also am faster than walking, hence the consequent is true. Thus, the counterfactual (3) is true. However, s'' is the closest world where the the antecedent of (4) is true, but there the puncture causes delay, so the consequent is false. Thus, the counterfactual (4) is false.

We now see how this idea is formalized.

6.2 Formal logic: counterfactuals

6.2.1 Formal semantics

Following the template for a state-based semantics, the formal semantics for counterfactuals is given as follows. After stating the formal definition, we explain it.

Definition 6.2. First, a similarity model M is a structure (S, R, I) where

- S is a nonempty set (the *state space* or *set of worlds*)
- R is a ternary relation on S (i.e., $R \subseteq S \times S \times S$)
- $I : S \times P \to \{0, 1\}$ is a function (interpretation)

satisfying (where, for $x \in S$, we define $A_x := \{y \in S : \exists z(Rxyz)\}$),

$$\forall x \in S\ \forall y \in A_x : Rxyy \qquad \text{(Reflexivity)}$$
$$\forall x \in S\ \forall y, z, w \in A_x : (Rxyz \text{ and } Rxzw) \Rightarrow Rxyw \qquad \text{(Transitivity)}$$
$$\forall x \in S\ \forall y, z \in A_x : (Rxyz \text{ and } Rxzy) \Rightarrow y = z \qquad \text{(Antisymmetry)}$$

Other versions instead require irreflexivity ($\not Rxyy$) and/or don't require antisymmetry. This doesn't affect the logic, but the setting of partial orders makes things easier to state (e.g., the limit assumption below).

We also write $y \leqslant_x z$ for $Rxyz$ and say 'y is more (or equally) *similar* to x than z'. Thus, the above conditions just say that each $\leqslant_x$ is a partial order on A_x. The worlds in A_x are also called the worlds *accessible* to x. The structure (S, R) is a *similarity frame* (i.e., a similarity models without an interpretation).

Second, we define when, in a model M, a formula φ is true at a world s (in which case we also say that s is a φ-*world*).

Only the last clause is new.

- $M, s \vDash p$ iff $I(s, p) = 1$.
- $M, s \vDash \top$ always, and $M, s \vDash \bot$ never.
- $M, s \vDash \neg\varphi$ iff $M, s \nvDash \varphi$.
- $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$.
- $M, s \vDash \varphi \vee \psi$ iff $M, s \vDash \varphi$ or $M, s \vDash \psi$.
- $M, s \vDash \varphi \to \psi$ iff, for all $x \in A_s$ with $M, x \vDash \varphi$, the following holds:

  there is some $y \in A_s$ with $Rsyx$ and $M, y \vDash \varphi$ such that, for any $z \in A_s$, if $Rszy$ and $M, z \vDash \varphi$, then $M, z \vDash \psi$.

  We explain below how this clause expresses the informal idea $(*)$.

And $\leftrightarrow$ is treated, as usual, as abbreviation for $(\varphi \to \psi) \wedge (\psi \to \varphi)$. Here we use the standard conditional symbol '$\to$' for the counterfactual. Other common symbols for the counterfactual are $>, \boxright, \Rightarrow, \rightsquigarrow$. If needed, we write $\varphi \supset \psi$ as an abbreviation for the material conditional $\neg\varphi \vee \psi$, and $\varphi \subset\!\supset \psi$ as abbreviation for the material biconditional. We also write $[\![\varphi]\!] = \{s \in S : s \vDash \varphi\}$ for the truth-set of φ in a model.

Third, counterfactual consequence $\Gamma \vDash_{\mathsf{CF}} \varphi$ is defined as expected: For all similarity models M and states s, if $M, s \vDash \psi$ for every $\psi \in \Gamma$, then $M, s \vDash \varphi$.

Let's motivate the formal definitions. First, the idea behind A_x: Intuitively, for the worlds in A_x (the accessible ones) it makes sense to think that x might have been one of them, while the worlds not in A_x (the inaccessible ones) are so dissimilar to x that x could not have been one of them.

Second, how does the clause for $\to$ express the intuitive idea $(*)$ that the closest φ-worlds are ψ worlds? It says that for all φ-worlds x accessible from s, there is a φ-world y more or equally similar to s than x such that it and all more similar φ-worlds z are also ψ-worlds. The reason for this complicated phrasing is that there might be φ-worlds that get ever closer

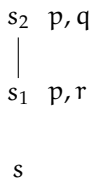

Figure 6.1: A similarity model violating strengthening the antecedent.

to s, so, strictly speaking, we cannot talk about 'the closest ones'. Instead, this is formalized as: there is a 'threshold' y such that anything closer to s than y counts as 'closest'—hence if these worlds are φ-worlds, they should be ψ-worlds.

If there are no chains of ever closer similarity, this complication can be neglected and we get a more straightforward formalization of $(*)$, as the following definition and exercise shows.

Definition 6.3. A similarity frame $M = (S, R)$ (and any model built over this frame) is said to satisfy the *limit assumption*, if, for all $s \in S$, the order $\leqslant_s$ is well-founded, i.e.,

> every nonempty subset $B \subseteq A_s$ has a minimal element, i.e., there is $m \in B$ such that, for all $b \in B$, if $b \leqslant_s m$, then $b = m$.

In other words, there are no infinitely descending chains $s_1 \geqslant_s s_2 \geqslant_s \ldots$ of distinct worlds that are increasingly more similar to s.

Exercise 6.4. Let $M = (S, R, I)$ be a similarity model satisfying the limit assumption. Show that the following are equivalent for any $s \in S$ and formulas φ and ψ:

1. $M, s \vDash \varphi \rightarrow \psi$ (as in definition 6.2)
2. For all $x \in S$, if x is a φ-world closest to s (i.e., x is a $\leqslant_s$-minimal element of $A_s \cap [\![\varphi]\!]$), then $x \vDash \psi$.

The direction (1)⇒(2) holds for any similarity model.

Before discussing the plausibility of the limit assumption, it's high time for a concrete example: the following shows that this formal semantics indeed violates the strengthening the antecedent.

Example 6.5. Consider the similarity model depicted in figure 6.1. It consists of:

- $S = \{s, s_1, s_2\}$
- $Rxyz$ iff $x = s$ and $y = s_i$ and $z = s_j$ for $i, j \in \{1, 2\}$ and $i \leqslant j$
- I maps to 1 the pairs $(s_1, p), (s_1, r), (s_2, p), (s_2, q)$ and maps all other inputs to 0.

This indeed defines a similarity model: First, note that $A_s = \{s_1, s_2\}$ and $A_{s_1} = \emptyset = A_{s_2}$. Then the conditions are easily seen to be satisfied: They hold trivially for $x \in \{s_1, s_2\}$ since both A_{s_1} and A_{s_2} are empty. So we need to check them for $x = s$. That's straightforward.

Note that, since S is finite, it cannot contain infinite decreasing chains, so we can use the clause for the counterfactual $\rightarrow$ under the limit assumption.

We see that $s \vDash p \rightarrow r$. Indeed, $[\![p]\!] \cap A_s = \{s_1, s_2\}$, so the only minimal element is s_1, and for it we have $s_1 \vDash r$.

And we see that $s \nvDash p \wedge q \rightarrow r$. Indeed, $[\![p \wedge q]\!] \cap A_s = \{s_2\}$, so a minimal element is s_2 (the only one), but for it we don't have $s_2 \vDash r$. ⌟

The limit assumption is debated. On the one hand, a standard example against it is, e.g., a stick which is exactly as long as a door wide: Consider the counterfactual 'if the stick were a little longer, it wouldn't fit through the door'. Among others, the antecedent makes us consider, for each n, the world s_n which is like the actual world except that the stick is $\frac{1}{n}$ cm longer. Arguably these worlds get ever closer to the actual world s, thus violating the limit assumption. On the other hand, it has been argued that, for all practical purposes, the context fixes a threshold beyond which any further difference doesn't matter to the context: in this case, say, a length difference < 0.1 mm doesn't matter.

See e.g. D. K. Lewis (1973, pp. 20–21).

See e.g. R. Stalnaker (1984, p. 141).

The logical difference that the limit assumption makes is, as exercise 6.c below shows, that it fails compactness. And, from a logical point of view, a failure of compactness is undesirable since it means that 'consistency' is not a 'finitary concept': it could be that a set of sentences has no model, but we will never get to notice that because all finite subsets (i.e., those that we can ever 'observe') do have a model. However, for validity, the limit assumption doesn't make a difference: a sentence is true at all states of all similarity models iff it is true it is true at all states of all similarity models satisfying the limit assumption. This is because similarity models have the so-called *finite model property* (as, e.g., the completeness proof of Burgess (1981) shows). This means that if a sentence fails on a similarity model, there also is a finite similarity model on which it fails, and those satisfy the limit assumption trivially.

Though, in another logical sense, compactness (of first-order logic) also is undesirable because it brings about, e.g., nonstandard models of Peano arithmetic PA *(since, for a new constant symbol* c*, the theory* $\mathsf{PA} \cup \{n \leqslant c : n \in \mathbb{N}\}$ *must have a model).*

Finally, a comment on alternative semantics: The present *comparative similarity semantics* is a prominent one, but there also are others.

1. *Sphere semantics*: Here the idea is to replace the relation R by a function $\$$ assigning each world s a collection $\$_s$ of subsets of S which intuitively are the spheres of similarity. So if $U \subseteq V$ are two such spheres, the elements of U are, up to a certain degree, similar to s, and the elements of V are, up to a more relaxed degree, similar to s. Thus, $s \vDash \varphi \to \psi$ iff there is a sphere $U \in \$_s$ with $\emptyset \neq [\![\varphi]\!] \cap U \subseteq [\![\psi]\!]$ (or no φ-world belongs to any sphere in $\$_s$).

   E.g. D. K. Lewis (1973, sec. 1.3).

2. *Selection function semantics*: Here the idea is to replace the relation R by a function $f : S \times \mathcal{P}(S) \to \mathcal{P}(S)$ that maps a pair $(s, [\![\varphi]\!])$ to a set of worlds $f(s, [\![\varphi]\!])$ which intuitively are the φ-worlds closest to s. So $s \vDash \varphi \to \psi$ iff $f(s, [\![\varphi]\!]) \subseteq [\![\psi]\!]$.

   A difference to the comparative similarity semantics: there, closeness of states is judged based on the states alone (language independently), while on the selection function semantics this is done relative to a proposition (language dependently). Exercise: Think about whether this is desirable.

   E.g. Starr (2019, sec. 2.3).

3. *Conditional semantics*: Here the idea is to replace the ternary relation R by a set of binary relations R_φ on the state space S, one for each sentence φ of our language $\mathcal{L}$. Intuitively, given a state s and a formula φ, the states x with $sR_\varphi x$ are the closest states making φ true. So So $s \vDash \varphi \to \psi$ iff, for all $x \in S$, if $sR_\varphi x$, then $x \vDash \psi$.

   The selection function is sometimes also formulated with a function mapping a state and a sentence (instead of a proposition) to a set of worlds. Exercise: Convince yourself that this is basically the same as the present conditional semantics.

   E.g. Priest (2008, ch. 5).

D. Lewis (1971, ch. 2) surveys these (and further) semantics and their relationships. Some are equivalent, some only under additional assumptions (e.g. on the selection function). Also see Schlechta and Makinson (1994) for a discussion (and a deep result) about when the similarity relation can be spelled out with a distance function (aka metric): i.e., $x \leqslant_s y$ iff the distance from s to x is smaller than the distance form s to y.

The relational semantics fits most neatly into the template of state-based semantics, and it strikes a good balance between simplicity and intuitiveness. Moreover, it generalizes well: by philosophically interpreting the relation R in other ways than just comparative similarity, we get logics for

Logic	$Rxyz$ interpreted as	$s \vDash \varphi \to \psi$ means
Counterfactuals	state y is more similar to state x than z is	If φ were the case, ψ would have been the case
Non-monotonic logic	In state x, y is more likely than z	If φ, then usually ψ
Belief revision	in belief state x, belief state y is more plausible than belief state z	After revision by φ, it is believed that ψ
Conditional obligation	in state x, the state y is preferable to state z	Given φ, it is obligatory that ψ

Figure 6.2: Interpretation of the ternary relation.

these interpretations. Examples are in figure 6.2 In the next chapter, we explore the non-monotonic logic interpretation.

Summarized, e.g., by Veltman (2006).

Finally, in philosophically discussions, the states are often taken as possible worlds. One should keep in mind, though, that this is an idealization. In linguistic practice, other parameters play an important role for conditionals, too, like time (see e.g. Khoo 2015). Though, toward a remedy, we can then take a state to be a pair (w, t) of a possible world and a point in time t.

6.2.2 Correspondence theory

Given a state-based semantics, an important question is: which properties of the models of the semantics correspond to which sentences of the logic? The reason is that it links up intuitive properties of the semantics with principles of the logic—often in non-obvious ways. This is helpful because both—properties and principles—can be useful guides in determining what 'the right' logic should be. Here is an example. To state it, we use some common terminology: A formula φ is *valid on* a similarity frame (S, R) if, for every interpretation I on (S, R) and every state $s \in S$, we have $(S, R, I), s \vDash \varphi$.

That's especially well explored in modal logic.

Exercise 6.6. Show that, for a similarity frame (S, R), the following are equivalent:

1. The formula $(\varphi \to \psi) \supset (\varphi \supset \psi)$ is valid on (S, R).
2. For all $s \in S$, s is a $\leqslant_s$-minimal element of A_s, i.e., $s \in A_s$ and, if $x \in A_s$ with $x \leqslant_s s$, then $x = s$.

Or, to be precise, every substitution instance of the formula is valid on (S, R).

The formula in (1) is called *modus ponens for* $\to$ (MP$\to$). And the condition (2) on frames is called *weak centering*.

Here both sides seem plausible: Semantically, it makes sense that there should be no world x that is strictly more similar to s than s itself. Logically, it makes sense that a counterfactual conditional connection (which quantifies about many worlds) between two sentences should be stronger than a material conditional connection (which quantifies only about the current world). After all, we said that typically a counterfactual $\varphi \to \psi$ is about a counterfactual situation φ, so typically φ is false at the current world, making the material conditional $\varphi \supset \psi$ true at the current world. This yields the intuitive reason for the correspondence: in the non-typical case that φ is true at the current world s (so it's not actually counterfactual), s should be a closest φ-world, hence it should also make ψ true, and thus the material conditional.

So this is a case of correspondence where both sides seem plausible, albeit for (prima facie) different reasons, so the correspondence says that these reasons are two sides of the same coin (as, on a second look, also became plausible). So this is a 'harmonic' example of correspondence, but there also are 'disharmonic' examples:

Exercise 6.7. Show that, for a similarity frame (S, R), the following are equivalent:

Again, rather every substitution instance of the formula is valid on (S, R).

1. The formula $\big((\varphi \to \psi) \wedge \neg(\varphi \to \neg\chi)\big) \supset \big((\varphi \wedge \chi) \to \psi\big)$ is valid on (S, R).

2. For all $s \in S$ and $x, y, z \in A_s$, if $x <_s z$, then either $x <_s y$ or $y <_s z$.

Note that $a <_s b$ *is* defined *to mean* $a \leqslant_s b$ *and* $a \neq b$.

The formula in (1) is called *rational monotonicity* (RM) or *strengthening with a possibility* (ASP). And the condition (2) on frames is called *almost connected*.

What does rational monotonicity say? It says that the antecedent of a counterfactual $\varphi \to \psi$ may be strengthened with χ as long as the antecedent doesn't exclude the possibility that χ. In other words, if in the closest counterfactual situations where φ is true, ψ is always true and χ can be true, also the closest counterfactual situations where both φ and χ are true, ψ is still always true. So this sharpens the idea that strengthening the antecedent shouldn't be valid: namely, by specifying conditions where strengthening the antecedent is still okay.

While this is, at least superficially, a plausible logical idea, the corresponding semantic idea of almost connectedness is a rather strong assumption about similarity: it says that if a world x is strictly more similar to s

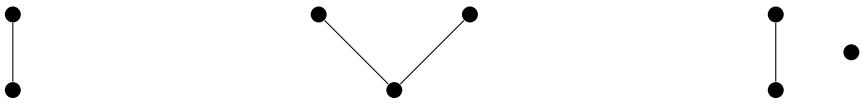

Figure 6.3: From left-to-right, a connected order, an almost connected but not connected order, and a not even almost connected order.

than a world z, then there cannot be a world y that is incomparable to both x and z in similarity to s. Note that this is weaker than the usual connectedness (aka totality aka linearity) assumption: that, for all x and y either $x \leqslant y$ or $y \leqslant x$: it's readily checked that this implies almost connectedness (justifiying the terminology), but not conversely as figure 6.3 shows.

Although almost connectedness for similarity is not very plausible, it is difficult to come up with natural language counterexamples to rational monotonicity. And, in fact, many logics of counterfactuals use it as an axiom.

6.2.3 Proof system

For completeness, and for those interested, we state a sound and complete proof system for counterfactuals. (We won't comment further on it here, but we'll discuss a very similar system in the next chapter.) The system is known as P (also as B). It's often regarded as the basic system for conditional logics (i.e., the analogue of K for modal logic).

Definition 6.8. The proof system P is described as follows. Axioms:

1. $\vdash \varphi$ if φ has the form of a Boolean classical tautology	(Taut)	*Tautologies*
2. $\vdash \varphi \to \varphi$	(CI)	*Conditional identity*
3. $\vdash \big((\varphi \to \psi) \wedge (\varphi \to \chi)\big) \supset \big(\varphi \to (\psi \wedge \chi)\big)$	(CC)	*Conj. of consequents*
4. $\vdash (\varphi \to \psi) \supset \big(\varphi \to (\psi \vee \chi)\big)$	(CW)	*Weakening the consequent*
5. $\vdash \big((\varphi \to \psi) \wedge (\varphi \to \chi)\big) \supset \big((\varphi \wedge \psi) \to \chi\big)$	(ASC)	*Strengthen Antec. with a Consequent*
6. $\vdash \big((\varphi \to \chi) \wedge (\psi \to \chi)\big) \supset \big((\varphi \vee \psi) \to \chi\big)$	(AD)	*Disjunction of Antec.*

Rules:

7. If $\vdash \varphi$ and $\vdash \varphi \supset \psi$, then $\vdash \psi$.	(MP$\supset$)	*Modus ponens for* $\supset$
8. If $\vdash \varphi \subset\!\supset \psi$, then $\vdash (\varphi \to \chi) \subset\!\supset (\psi \to \chi)$	(REA)	*Replacement of Equivalent Antecedents*
9. If $\vdash \varphi \subset\!\supset \psi$, then $\vdash (\chi \to \varphi) \subset\!\supset (\chi \to \psi)$	(REC)	*Replacement of Equivalent consequents*

The system P was proven to be sound and complete with respect to the relational semantics by Burgess (1981) and Veltman (1985), but the proofs are quite complicated.

6.3 Assessment

6.3.1 On the notion of similarity

The crucial question, now, is how well the formal semantics captures philosophical intuitions—and this hinges on the notion of similarity. An early argument to the contrary is as follows:

> "Consider a man – call him Jones – who is possessed of the following dispositions as regards wearing his hat. Bad weather invariably induces him to wear a hat. Fine weather, on the other hand, affects him neither way: on fine days he puts his hat on or leaves it on the peg, completely at random. Suppose, moreover, that actually the weather is bad, so Jones *is* wearing his hat" (Tichý 1976, pp. 271–272).

Consider the counterfactual "If it the weather had been fine, Jones would have been wearing his hat".

Intuitively, you probably would say this sentence is false. But a natural modeling in the relational semantics seems to make it true: In the actual world s it is raining and Jones wears a hat. We consider two worlds s_1 and s_2 that are like the actual world except that the weather is fine and in s_1 Jones wears his hat but not so in s_2. So s_1 is more similar to s since they still agree on Jones wearing his hat. So s_1 is the closest antecedent world, and the consequent is true, so the counterfactual is true.

6.4 Exercises

From Veltman (2006).

The first two exercises are on correspondence theory:

Exercise 6.a (Practice)**.** Do exercise 6.6.

Exercise 6.b (Problem)**.** Do exercise 6.7.

The next exercise is, in a sense, still about correspondence theory: but now characterizing a frame property (namely the limit assumption) with infinitely many sentences. The conclusion then is that the logical difference that the limit assumption makes is that compactness fails.

Exercise 6.c (Problem)**.** *From Veltman (2006).*

1. Let Γ contain, for each $n = 1, 2, 3, \ldots$, the following sentences:

$$\varphi_n := (p_1 \vee \ldots \vee p_{n+1}) \to \neg(p_1 \vee \ldots \vee p_n)$$
$$\psi_n := \neg\big((p_1 \vee \ldots \vee p_{n+1}) \to (p_1 \vee \ldots \vee p_n)\big).$$

   Let (S, R) be a similarity frame. Show: there is *no* interpretation I on (S, R) and state s such that $(S, R, I), s \vDash \chi$ for each $\chi \in \Gamma$ iff (S, R) satisfies the limit assumption.

2. Conclude that the logic of similarity models satisfying the limit assumption is not compact: i.e., show that the infinite set Γ is such that each finite subset Γ_0 is satisfied in some state s of a similarity model M satisfying the limit assumption (i.e., $M, s \vDash \psi$ for each $\psi \in \Gamma_0$), but all of Γ cannot be satisfied at a state of a similarity model satisfying the limit assumption.

Exercise 6.d (Philosophical)**.** You can choose to write about one of the following topics:

1. Consider the objections to a (or Lewis') similarity analysis of counterfactuals. For example, Tichý's (section 6.3.1). Or the one of Fine (1975, p. 452) that the analysis makes true the intuitively false counterfactual (2) 'If Oswald hadn't shot Kennedy, then someone else would have': "on the grounds that the consequences of supposing that someone else shot Kennedy would make less difference to the world than those of supposing that Kennedy was not shot after all."

   Do you think these are decisive arguments against similarity-based analyses of counterfactuals (if so, what would be the general, not example-based argument)? Or do you have a reply—e.g., that similarity has to be spelled out more carefully? If so, does the reformulation yield other problems?

2. What to philosophically make of the correspondence between rational monotonicity and almost connectedness (exercise 6.7)? Do you have arguments for rational monotonicity and hence want to accept almost connectedness? Do you think almost connectedness is plausible after all (maybe on a certain conception of similarity)?

3. It has been argued (e.g. Fine 2017, p. 571) that the sentential context of counterfactuals is *hyperintensional*: The following counterfactuals

intuitively differ in truth-value even though their antecedents are necessarily equivalent

(a) If Sue were to take the pill, then she would live.

(b) If Sue were to take the pill or to take the pill and the cyanide, then she would live.

The good ol' p vs $p \vee (p \wedge q)$ showing up again.

Similarly, counterfactuals with different but impossible antecedents (aka *counterpossibles*) formally always come out true but intuitively may differ in truth-value (e.g. Williamson 2018). Is this failure in distinguishability a shortcoming or can it be defended?

Again, focus on a concrete aspect of the question that you find interesting, provide clear and careful arguments, and consult the cited literature if you look for some inspiration.

6.5 Notes

Priest (2008, ch. 5), Veltman (2006), and Burgess (1981).

7 Non-monotonic logics

Non-monotonic logics describe 'common-sense' or 'defeasible' reasoning: when we think about birds, we conclude that, usually, they can fly—even though this can be defeated by the exception of, say, penguins. A semantics for such logics can be given, essentially, by a reinterpretation of the relational semantics for counterfactuals that we've seen in the last chapter. Instead of ordering the worlds by similarity, we now think of the order as describing when one world is more usual than another. Thus, a defeasible reasoning step $\varphi \mathrel{|\!\sim} \psi$ is considered correct if the most usual φ-worlds also are ψ-worlds. We describe in detail the resulting semantics and provide the corresponding sound and complete laws of defeasible reasoning.

Key concepts
- Defeasible vs. deductive reasoning
- Non-monotonic vs. monotonic logic
- Plausible consequence
- Preferential consequence relations and system P
- KLM semantics in terms of preferential models
- Soundness and completeness
- Frame problem (in artificial intelligence)

7.1 Motivation

We've seen that counterfactuals don't validate strengthening of the antecedent: the truth of $\varphi \to \psi$ doesn't imply the truth of $\varphi \wedge \chi \to \psi$. This is a feature that is by far not specific to counterfactuals. It is very common—some might say defining—for conditionals representing 'everyday reasoning':

1. ✓ If it is a bird, then it can fly.
   × If it is a bird in Antarctica, then it can fly.
2. ✓ If you flip the switch, then the light will turn on.
   × If you flip the switch and there is a power outage, then the light will turn on.
3. ✓ If the street is wet, then it is raining.

$\times$ If the street is wet and covered by a roof, then it is raining.

This kind of reasoning is called *defeasible reasoning*: unlike deductive reasoning, defeasible reasoning allows for exceptions (defeating previous reasoning steps). The reasoner might still retract their conclusion in light of further information (e.g., retracting the conclusion that the bird can fly when learning that we might be dealing with penguins).

Logics that describe this defeasible reasoning are called *non-monotonic logics*, because an increase in assumptions might yield a decrease in conclusions. Just as with counterfactuals, this also means that most of the logics that we've seen cannot model defeasible reasoning because they are monotonic. In particular, not only counterfactual conditionals but also the above indicative 'common sense' conditionals need a more careful treatment than as a material (or other monotonic) conditional.

Since we talk about reasoning (correctly going from premises to a conclusion), it arguably is more natural to express these natural language conditionals as instances of a consequences relation like $\varphi \models \psi$ (meta-language), rather than as the truth of a conditional $\varphi \to \psi$ (object-language). So we think of the sentence 'if it is a bird, then it can fly' as a correct *contentful* rule of inference rather than expressing a true sentence.

'Contentful' in the sense that it is about birds and flying while a usual rule of inference like 'If $\varphi \wedge \psi$, then φ' is purely formal without content (which is taken as hallmark for logical inferences).

To some extent this is a matter of preferred terminology, but choosing consequence relations simplifies things: while conditionals can be nested (e.g., $\varphi \to (\varphi \to \psi)$), consequence relations cannot (it makes no sense to write $\varphi \models (\varphi \models \psi)$). Thus, we can work with a Boolean language (i.e., one that doesn't contain a conditional) because we don't need the conditional to express defeasible reasoning.

This is the common setup of non-monotonic logics: a Boolean language and a consequence relation aiming to capture defeasible reasoning. The consequence relation is then written as $\mathrel{|\!\sim}$ to stress that it doesn't have the usual properties we'd expect from a consequence relation $\models$ capturing deductive reasoning. So $\varphi \mathrel{|\!\sim} \psi$ is understood as

If φ is the case, then *usually* ψ is the case.

These statements $\varphi \mathrel{|\!\sim} \psi$ are also called *plausible consequence*.

The difference to $\models$ is this: Semantics for $\varphi \models \psi$ are centered around the idea (going back to Tarski) that in *all* situations (or models) where φ is true, also ψ is true. But for $\mathrel{|\!\sim}$ the idea is that only for certain situations (or models) where φ is true—namely, the *usual* ones—also ψ needs to be true. Instead of 'usual' one might also speak of *plausible*, *likely*, *normal*, or *typical*.

The task of a non-monotonic logic is to specify a defeasible consequence relation $\mathrel{|\!\sim}$. And with the just mentioned insight about the intuitive meaning of $\varphi \mathrel{|\!\sim} \psi$ as 'truth-preservation in the most usual situations' (instead of all situations), we're prompted to turn (back) to counterfactuals (i.e., think of $\varphi \mathrel{|\!\sim} \psi$ as $\varphi \rightarrow \psi$). Indeed, semantically a counterfactual expresses truth preservation in the most similar situations. So we might just reinterpret the formal relation R in the relational semantics for counterfactuals. Thus, instead of similiarity, we think of $Rsxy$ as x is more normal/likely/etc. than y (with respect to the base world s). Moreover, proof-theoretically, we can also look at principles for counterfactuals $\rightarrow$ and see whether they translate into laws for nonmonotonic consequence $\mathrel{|\!\sim}$. After all, we've already seen that, just like for counterfactuals, antecedent strengthening should also fail for plausible consequence.

This was one of the interpretations of R *from figure 6.2.*

We'll now do this formally in the next section. After that, we'll see why the invention of nonmonotonic logic was important to artificial intelligence.

7.2 Formal logic: KLM

There are many non-monotonic logics: logic programming (with negation as failure), default logic, autoepistemic logic, etc. (Strasser and Antonelli 2019). A general framework including these was provided by Kraus et al. (1990). This classic paper is often referred to by 'KLM' indicating the authors Kraus, Lehmann, and Magidor. We present a summarized version here.

Not the airline!

7.2.1 Proof theory: the system P

Abstractly speaking, we say a *consequence relation* $\mathrel{|\!\sim}$ is a relation between subsets Γ of sentences and sentences φ of some background language. Any logic that we have considered so far provides such a consequence relation (e.g., $\models_{CL}, \models_{IL}, \models_{CF}, \ldots$). But this abstract definition, of course, also includes many consequence relations not worth the name (e.g., those not containing $\{p\} \mathrel{|\!\sim} p$). So the task is to describe which are the 'good' ones. One might say that the consequence relation $\models_{CL}$ of classical logic is a good candidate for capturing the general laws of deductive reasoning. But here we're interested in defeasible reasoning. So the task for us now is to describe which consequence relations $\mathrel{|\!\sim}$ are good candidates for capturing the general laws of defeasible reasoning. A standard 'baseline' answer is given in the definition below.

In fact, for our purposes we can simplify the notion of a consequence relation: First, we can take the Boolean language $\mathcal{L}_{\text{bool}}$ as background language. Second, we restrict us to finite Γ (and, if needed, later extend it to infinite Γ assuming compactness). Third, we assume $\varphi_1, \ldots, \varphi_n \mathrel{|\!\sim} \varphi$ is (defined to be) the same as $\varphi_1 \wedge \ldots \wedge \varphi_n \mathrel{|\!\sim} \varphi$. Thus, we can take Γ to really be given by a single sentence. So we take a consequence relation $\mathrel{|\!\sim}$ to be a binary relation on $\mathcal{L}_{\text{bool}}$. (If we regard a plausible consequence $\varphi \mathrel{|\!\sim} \psi$ as really just a pair (φ, ψ) we can think of $\mathrel{|\!\sim}$ as a set of plausible consequences.)

Cf. the difference of distributive entailment and conjunctive entailment in truthmaker semantics (example 5.2).

Definition 7.1. A consequence relation $\mathrel{|\!\sim}$ is called *preferential* if it satisfies, for all $\varphi, \psi, \chi \in \mathcal{L}_{\text{bool}}$:

1. *Reflexivity*: $\varphi \mathrel{|\!\sim} \varphi$.
2. *Left logical equivalence*: If φ and ψ are classically equivalent and $\varphi \mathrel{|\!\sim} \chi$, then $\psi \mathrel{|\!\sim} \chi$.
3. *Right weakening*: If φ classically entails ψ and $\chi \mathrel{|\!\sim} \varphi$, then $\chi \mathrel{|\!\sim} \psi$.
4. *Cut*: If $\varphi \wedge \psi \mathrel{|\!\sim} \chi$ and $\varphi \mathrel{|\!\sim} \psi$, then $\varphi \mathrel{|\!\sim} \chi$.
5. *Cautious monotonicity*: If $\varphi \mathrel{|\!\sim} \psi$ and $\varphi \mathrel{|\!\sim} \chi$, then $\varphi \wedge \psi \mathrel{|\!\sim} \chi$.
6. *Or*: If $\varphi \mathrel{|\!\sim} \chi$ and $\psi \mathrel{|\!\sim} \chi$, then $\varphi \vee \psi \mathrel{|\!\sim} \chi$.

This collection of rules is known as *system* P.

Four comments: First, the weaker system obtained from P by removing the or-rule (rule 6) is known as C. It has been argued (e.g., by Gabbay) to collect the rockbottom properties a consequence relation should have. Its advantage is that it can be formulated independently of the background language (which, for us, is $\mathcal{L}_{\text{bool}}$): then we don't need the rule for $\vee$ anymore and we replace occurrences of $\wedge$ like $\varphi \wedge \psi \mathrel{|\!\sim} \chi$ by $\varphi, \psi \mathrel{|\!\sim} \chi$ (and hence work with the more general definition of a consequence relation).

Though one needs to be careful with 'classical entailment/tautology' occurring in the rules; and one may also contemplate taking another logic than classical logic for this.

Second, another well-known system—strictly in between C and P in terms of logical strength—is CL (not to be confused with classical logic) obtained from C by adding the rule:

Loop: If $\varphi_0 \mathrel{|\!\sim} \varphi_1, \ldots, \varphi_{n-1} \mathrel{|\!\sim} \varphi_n$, and $\varphi_n \mathrel{|\!\sim} \varphi_0$, then $\varphi_0 \mathrel{|\!\sim} \varphi_n$.

But here we'll focus on P.

Third, as suggested in the motivation section, the resemblance between counterfactuals and plausible consequences shows up in the axioms: We've stated a sound and complete proof system for counterfactuals in

section 6.2.3. And it essentially translates into the system P here by replacing $\rightarrow$ with $\mathrel{\sim}$: For example, the conditional identity axiom $\varphi \rightarrow \varphi$ becomes the reflexivity rule. The strengthening the antecedent with a consequent axiom $\big((\varphi \rightarrow \psi) \wedge (\varphi \rightarrow \chi)\big) \supset \big((\varphi \wedge \psi) \rightarrow \chi\big)$ becomes the cautious monotonicity rule. And so on. (Though, there are some details that don't fit directly: e.g., the weakening of the consequent axiom $(\varphi \rightarrow \psi) \supset \big(\varphi \rightarrow (\psi \vee \chi)\big)$ would become 'If $\varphi \mathrel{\sim} \psi$, then $\varphi \mathrel{\sim} \psi \vee \chi$', but this is equivalent to the right weaking rule above.) One also says that P here is the *flat fragment* of the system for counterfactuals, since it only considers non-nested counterfactuals.

Fourth, the kind of proof system that P is an instance of is know as a *sequent calculus*. While proof systems like natural deduction or Hilbert systems provide rules and axioms to derive sentences φ, sequent calculi are 'a level higher' and provide rules and axioms to derive sequents $\varphi \Rightarrow \psi$. These sequents are meant to say that φ implies ψ in the intended sense. So a sequent calculus for classical logic provides rules and axioms for deriving classical consequences $\varphi \Rightarrow \psi$. Here the sequent is written $\varphi \mathrel{\sim} \psi$ and is meant to say that φ has the plausible consequence ψ. And system P provides rules and axioms to derive plausible consequences from other ones.

7.2.2 Semantics: preferential models

The semantics will be reminiscent of the relational semantics for counterfactuals, but also somewhat different. In particular, it will not be a standard state-based semantics (as specified in our template). This is because the relevant notion will be when a plausible consequence $\varphi \mathrel{\sim} \psi$ is true 'globally' at a model M, instead of providing a notion of 'local truth' at a state.

Definition 7.2. A *preferential model* M is a structure (S, R, I) where

- S is a set (state space)
- R is a binary relation on S (preference) that is irreflexive ($\forall x \in S : x \not R x$) and transitive ($\forall x, y, z \in S : xRy \;\&\; yRz \Rightarrow xRz$)
- $I : S \times P \rightarrow \{0, 1\}$ is a function (interpretation) extended to all formulas of $\mathcal{L}_{\text{bool}}$ in the classical way (i.e., $I(s, \neg\varphi) = 1$ iff $I(s, \varphi) = 0$; and $I(s, \varphi \wedge \psi) = 1$ iff $I(s, \varphi) = 1$ and $I(s, \psi) = 1$; etc.)

Here $a \not R b$ means that it is not the case that aRb

such that the so-called *smoothness condition* is satisfied: To define it, first say that a state $s \in S$ is R-*minimal in* a subset $A \subseteq S$ if $s \in A$ and $\forall a \in A : a \not R s$;

and, second, say that a subset $A \subseteq S$ is *smooth* if, for all $a \in A$, either a is R-minimal in A or there is $a' \in A$ with $a'Ra$ and a' is R-minimal in A. Then:

- Smoothness condition: For all formulas $\varphi \in \mathcal{L}_{\text{bool}}$, the truth-set $[\![\varphi]\!] := \{s \in S : I(s, \varphi) = 1\}$ is smooth.

The states of M act like classical states: $s \vDash \varphi$ iff $I(s, \varphi) = 1$. The interesting bit is the consequence relation: We say M *makes true* or *validates* the plausible consequence $\varphi \mathrel{|\!\sim} \psi$ (and write $\varphi \mathrel{|\!\sim}_M \psi$) if for all R-minimal elements s of $[\![\varphi]\!]$, we have $s \in [\![\psi]\!]$ (i.e., $I(s, \psi) = 1$). We say $\mathrel{|\!\sim}_M$ is the preferential consequence relation *defined by* M. (The soundness theorem below shows that it is indeed a preferential consequence in the sense of definition 7.1.)

We say a set K of plausible consequences *(preferentially) entails* a plausible consequence $\varphi \mathrel{|\!\sim} \psi$ if, for all preferential models M, if $K \subseteq \mathrel{|\!\sim}_M$ (i.e., every plausible consequence of K is validated by M), then $\varphi \mathrel{|\!\sim}_M \psi$.

We can think of K as a knowledge base (hence the letter 'K').

Again as suggested in the motivation section, we see the similarity to counterfactuals also on the semantic side. A preferential model $M = (S, R, I)$ is much like a similarity model of the semantics for counterfactuals:

First, we now require irreflexivity (instead of reflexivity) and antisymmetry then follows from transitivity. (To see this: Write $x\underline{R}y$ for 'xRy or $x = y$'. Then $x\underline{R}y$ and $y\underline{R}x$ implies $x = y$, because if $x\underline{R}y$ and $y\underline{R}x$ but $x \neq y$, we must have xRy and yRx, which, by transitivity, implies xRx, which cannot be.) But we already said that the relational semantics for counterfactuals could equivalently be done in this setting as well.

Second, the relation R describes 'preference' or 'normality': xRy is intended to mean that world x is more normal (or preferable) than world y. This is now not relative to a base world s, hence R is binary rather than ternary. Intuitively, one can express this by saying that the choice of base world s is irrelevant: i.e., take R as a ternary relation, but require that two base worlds give rise to the same order (aka absoluteness: for all $s, s' \in S$, we have $\leqslant_s = \leqslant_{s'}$) and that we don't need to separately keep track of accessibility (aka universality: for all $s \in S$, we have $A_s = S$). To formally obtain a preferential model this way, we would need to have started with an irreflexive and transitive ternary relation R, though.

Third, the smoothness assumption is very similar to the limit assumption for similarity models. However, it is not a purely frame-based condition, but also requires the interpretation (to define the truth-sets). On the other hand, it then requires minimal elements only for nonempty truth-set and

not for any nonempty subset. (The limit assumption for counterfactuals is sometimes also formulated dependent on the interpretation and not purely frame-based.)

Fourth, in this setup, the truth of a counterfactual $\varphi \to \psi$ is essentially the truth of the plausible consequence $\varphi \mathrel{|\!\sim} \psi$: the minimal φ-worlds are ψ-worlds. The difference is that for counterfactuals the intended interpretation of 'minimal' is 'most similar' (or 'closest'), while for plausible consequences it is 'most normal'. And for counterfactuals we specify their truth relative to a state, while for plausible consequences we explained that any base world is as good as any other, so we can just omit them and only say when the whole model makes true the plausible consequence.

Exercise 7.3 (KLM's penguin triangle)**.** Let K contain the following three plausible consequences (where p stands for penguin, b for bird, and f for fly):

$$\mathsf{p} \mathrel{|\!\sim} \mathsf{b} \qquad\qquad \mathsf{p} \mathrel{|\!\sim} \neg \mathsf{f} \qquad\qquad \mathsf{b} \mathrel{|\!\sim} \mathsf{f}$$

1. Show that $\mathsf{p} \mathrel{|\!\sim} \mathsf{f}$ is *not* entailed by K.
2. Show $\mathsf{p} \wedge \mathsf{b} \mathrel{|\!\sim} \neg \mathsf{f}$ is entailed by K.

7.2.3 Soundness and completeness

We now see that the proof-theoretic description of plausible consequence (the laws for defeasible reasoning) aligns exactly with the semantic one (plausible consequence as truth-preservation in the most normal worlds).

Theorem 7.4. *1. Soundness: For every preferential model M, the consequence relation $\mathrel{|\!\sim}_{\mathsf{M}}$ is preferential.*

2. Completeness: Every preferential consequence relation $\mathrel{|\!\sim}$ is given as $\mathrel{|\!\sim}_{\mathsf{M}}$ of some preferential model M.

The proof is due to Kraus et al. (1990), using techniques of Veltman (1985). We sketch it at the end of this subsection, but first a corollary in the more typical soundness and completeness form.

Corollary 7.5. *For a set K of plausible consequences and a plausible consequence $\varphi \mathrel{|\!\sim} \psi$, the following are equivalent:*

1. K preferentially entails $\varphi \mathrel{|\!\sim} \psi$ (i.e., for all preferential models M, if $\mathsf{K} \subseteq \mathrel{|\!\sim}_{\mathsf{M}}$, then $\varphi \mathrel{|\!\sim}_{\mathsf{M}} \psi$).

2. *Using the elements of* K *as additional axioms, one can derive, using the rules of system* P*, the plausible consequence* $\varphi \mathrel{|\!\sim} \psi$.

Proof sketch of corollary 7.5. For (2)⇒(1), use soundness (theorem 7.4 (1)). For (1)⇒(2), assume (2) is false. Then the smallest consequence relation $\mathrel{|\!\sim}$ closed under the rules of P and containing K is a preferential consequence relation with $\varphi \not\mathrel{|\!\sim} \psi$. By completeness (theorem 7.4 (2)), there is a preferential model M with $\mathrel{|\!\sim} = \mathrel{|\!\sim}_M$. So (1) fails. □

Proof sketch of theorem 7.4. Soundness is a matter of checking that the rules for a preferential consequence relation are satisfied. For completeness, let $\mathrel{|\!\sim}$ be a preferential consequence relation and construct $M = (S, R, I)$ as follows:

- S is the set of pairs (ν, φ) where $\nu : P \to \{0, 1\}$ is a classical valuation and $\varphi \in \mathcal{L}_{\text{bool}}$ a sentence such that, for all $\psi \in \mathcal{L}_{\text{bool}}$, if $\varphi \mathrel{|\!\sim} \psi$, then $\nu(\psi) = 1$.

  KLM then call ν a normal world for φ.

- $(\nu, \varphi) R (w, \psi)$ iff $\varphi \vee \psi \mathrel{|\!\sim} \varphi$ and $\nu \not\models \psi$
- $I\big((\nu, \varphi), \psi\big) := \nu(\psi)$.

One then checks that this indeed defines a preferential model and that $\mathrel{|\!\sim}_M = \mathrel{|\!\sim}$. □

Similar semantics can also be given for the weaker systems C and CL. They generalize the relation R to any binary relation respectively transitive ones. But the interpretation $I(s, \cdot)$ isn't anymore a single classical valuation, but actually a set of classical valuations.

7.3 Assessment

Non-monotonic logics have been developed—at least in part—to provide a solution to the *frame problem* in artificial intelligence (Shanahan 2016). It is usually put as the problem of logically representing the effects of actions without having to explicitly represent all the intuitive non-effects.

For example, we might build an 'AI' to automatically adjust the light and temperature in a room. We might build in the rules (all informal in the following)

1. If the daylight sensor is low, turn on the light.
2. If the temperature is low, turn on the heating.

Now assume that first the sensor is low, and the AI turns on the light. A little later, the temperature is low, and the AI turns on the heating. Intuitively, we would think that the light is still on, i.e., the action of turning on the heating didn't interfere with the light being on. But on a common formalization based on classical logic (e.g., situation calculus), this doesn't follow and we would need to add this as an additional rule (these rules are called *frame axioms*).

However, it would be inefficient to explicitly represent each such frame axiom for almost any pair of two actions (since most actions usually are independent of each other). It would be more principled to let the AI reason according to common sense (i.e., defeasibly) rather than classically (i.e., deductively). That's the solution to the frame problem provided by non-monotonic logics: change the background logic from classical logic to an appropriate non-monotonic logic. Then the frame axioms need not explicitly be represented since, intuitively, they are true: If the heating is turned on, then, in the most normal world, the light will continue to be on.

Moving to a non-monotonic background logic also solves another issue which is closely connected to the frame problem: If we understand rules like the frame axioms classically, they cannot allow for exceptions. For example, consider the frame axiom

> If the heating is turned on, the light will continue to be on.

It can also be applied when turning on the electricity-consuming heating caused a fuse to go off, yielding the false conclusion that the light is on. (This is just the fact that classical logic is monotonic.) So we better add that exception to the rule:

> If the heating is turned on and doesn't consume too much energy, the light will continue to be on.

But now we're on a slippery slope. There is an infinite number of potential exceptions to this rule: breaking light bulbs, lightning strikes, alien attacks, etc. But we cannot possibly explicitly represent all these exceptions. Non-monotonic logics solve this issue: they assume that usually there are no exceptions, but there may be some in non-usual worlds.

For more on the use of non-monotonic logics (specifically logic programming) in modeling actual human reasoning, see Stenning and van Lambalgen (2008).

7.4 Exercises

Exercise 7.a. For a consequence relation $\mathrel{|\!\sim}$, consider the *equivalence rule*:

($*$) If $\varphi \mathrel{|\!\sim} \psi$ and $\psi \mathrel{|\!\sim} \varphi$, then, if $\varphi \mathrel{|\!\sim} \chi$, also $\psi \mathrel{|\!\sim} \chi$.

Show both proof-theoretically and semantically that ($*$) is derivable in P. In other words:

1. Show that if $\mathrel{|\!\sim}$ is preferential, then it satisfies ($*$). (Use the properties from definition 7.1 that $\mathrel{|\!\sim}$ then has.)

2. Show that if M is a preferential model, then $\mathrel{|\!\sim}_M$ satisfies ($*$). (Use the semantic definition of $\mathrel{|\!\sim}_M$ from definition 7.2. In particular, don't use (1) and theorem 7.4 saying that $\mathrel{|\!\sim}_M$ is preferential.)

Philosophically reflect on the plausibility of this rule.

E.g., on the dynamical system interpretation of non-monotonic logics of Leitgeb (2005).

7.5 Notes

Kraus et al. (1990), Veltman (2006), and Strasser and Antonelli (2019).

8 Relevance logic

Relevance logics provide a conditional where the antecedent must be, in some sense, relevant to the consequence: so, unlike the material conditional, one should *not* have validities like $\varphi \to (\psi \to \psi)$ where the obtaining of φ is irrelevant to the obtaining of the triviality $\psi \to \psi$. In this chapter, we consider a version of the relational semantics for counterfactuals to provide such a 'relevant' semantics for conditionals. The intuitive interpretation of the ternary relation $Rxyz$ allows several interpretations but is along the lines of: if, in information state x, one adds the information of state y, one obtains only information that's already in z. And the clause for the conditional now also changes: instead of 'all closest antecedent-worlds also are consequent-worlds' one gets $x \vDash \varphi \to \psi$ iff whenever $Rxyz$ with $y \vDash \varphi$, also $z \vDash \psi$.

Key concepts •

TBA

8.1 Motivation

We've seen that counterfactual conditionals and indicative 'common sense' conditionals aren't monotone and hence cannot be described by the material conditional (or other monotone ones). But monotonicity isn't the only material validity that should be avoided to adhere to our intuitions about 'correctness' of conditionals.

The following sentences are validities for the material conditionals (and many other ones, too):

See, e.g., Priest (2008, sec. 1.9).

$$\varphi \to (\psi \to \psi)$$
$$\varphi \to (\psi \vee \neg\psi)$$
$$(\varphi \wedge \neg\varphi) \to \psi$$
$$\big(\varphi \to (\psi \to \chi)\big) \to \big((\varphi \to \psi) \to (\varphi \to \chi)\big)$$
$$(\varphi \to \psi) \vee (\psi \to \varphi)$$
$$\neg(\varphi \to \psi) \to \varphi$$

However, intuitively, it's not clear why they should be valid: For example,

regarding the first, what does φ have to do with the truth of the validity of $\psi \to \psi$? Regarding the last, plausibly, "It is not the case that, if there is a good god, then the prayers of evil people will be answered"; but why should that imply that there is a good god? (Priest 2008, p. 15). Exercise: think about the other sentences.

That's why these and similar sentences are called the *paradoxes of the material conditional*. And this is already the case for 'simple' indicative conditionals, and thus a wide-spread phenomenon.

Relevance logics have been developed to provide a conditional where the antecedent is relevant to the consequent, to exclude irrelevancies like the first three sentences above.

One might ask: do we really need a new logic for this? Two more specific versions of the question:

First, can't we use an existing logic? For example, counterfactuals? No, counterfactuals still validate the irrelevancy $\varphi \wedge \neg\varphi \to \psi$ because counterfactuals with impossible antecedents are always true on the relational semantics (cf. exercise 6.d (3) on counterpossibles). Intuitionistic logic, also wouldn't violate this but it violates, e.g., $(\varphi \to \psi) \vee (\psi \to \varphi)$ (see exercise 4.18 on the Gödel–Dummett logic where this is added as an additional axiom). The logic LP violates $\varphi \wedge \neg\varphi \to \psi$, but it would satisfy $\varphi \to (\psi \to \psi)$ (because this would be undesignated, i.e., 0, only if φ is 1 and ψ is 0, but $\psi \to \psi$ is, according to the strong Kleene truth-table always i or 1). Exercise: go through some more logics.

Second, should a logic at all respect these intuitions of relevance? Can't we just say that classical logic is fine and these irrelevancies are only a pragmatic weirdness, but not semantically wrong? Some relevant logicians might argue that this is more than just pragmatics, but one might also concede the point and say one still would like to develop a logic describing these pragmatics. After all, they are not completely ill-behaved but still seem to follow some rules—and it would be the task of the logic to capture and explain these rules.

Let's now see how to provide a semantics—based on a version of the ternary relation semantics for counterfactuals—to express this idea of relevance.

8.2 Formal logic: basic relevant logic

We specify the semantics following the state-based semantics template. It was invented by Routley and Meyer in 1973, and later got simplified by

Priest and Sylvan in 1992 and then got further simplified by Restall (1993). We say more on the intuitive interpretation, especially of the relations R and $*$, afterward.

Definition 8.1. A *Routley–Meyer model* M is a structure $(S, R, N, *, I)$ where

- S is a set (state space or *set of worlds*)
- R is a ternary relation on S (i.e., $R \subseteq S \times S \times S$)
- N is a unary relation on S (i.e., $N \subset S$), the elements of N are called *normal worlds* and the elements of $S \setminus N$ are the *non-normal worlds*
- $*$ is a function $S \to S$ with period two ($s^{**} = s$) called the *Routley Star* (and s^* is called the *twin* of s)
- $I : S \times P \to \{0, 1\}$ is a function (*interpretation*)

satisfying the *normality condition*

> For all $s, s_1, s_2 \in S$, if $s \in N$, then Rss_1s_2 iff $s_1 = s_2$.

This structure but without the interpretation is a *Routley–Meyer frame*.

Second, we recursively define when a state s makes true a formula φ ($M, s \vDash \varphi$)

- $s \vDash p$ iff $I(s, p) = 1$
- $s \vDash \neg\varphi$ iff $s^* \nvDash \varphi$
- $s \vDash \varphi \wedge \psi$ iff $s \vDash \varphi$ and $s \vDash \psi$
- $s \vDash \varphi \vee \psi$ iff $s \vDash \varphi$ or $s \vDash \psi$
- $s \vDash \varphi \to \psi$ iff for all $s_1, s_2 \in S$, if Rss_1s_2 and $s_1 \vDash \varphi$, then $s_2 \vDash \psi$.

Third, consequence is almost defined as usual, we now only restrict it to normal worlds: $\Gamma \vDash_B \varphi$ iff for all Routley–Meyer models $M = (S, R, N, *, I)$ and $s \in N$, if $M, s \vDash \psi$ for every $\psi \in \Gamma$, then $M, s \vDash \varphi$. We use the subscript B because the relevance logic which is sound and complete with respect to this semantics is known as the basic relevant logic B.

Four comments: First, this is an instance of a state-based semantics where we use more than just one relation: we use the unary relation N, the function $*$ (i.e., a special case of a binary relation), and the ternary relation R.

Second, as for the relational semantics for counterfactuals, the relation R is used to interpret the conditional. Though, the clause for the conditional

is quite different: it now says that if the base world s makes true the conditional $\varphi \to \psi$, then, for any accessible world s_1 that 'links' to another world s_2, this link ensures that the truth of φ at s_1 implies the truth of ψ at s_2. We say more on how to interpret R below.

Third, why the business of non-normal worlds? The idea is, that if we want to violate $\varphi \to (\psi \to \psi)$, we should expect a world s where $\psi \to \psi$ fails, but this must be a 'non-normal' world because no 'classical' or 'normal' world could violate such a triviality. Consequence then should still be truth-preservation at the normal worlds: intuitively, those are the 'serious' ones we care about, while the non-normal worlds are those we need to check for relevance. The hallmark for being a normal world is expressed by the normality condition: for normal worlds, the ternary relation R essentially is a binary relation sRs' making normal worlds act much like possible worlds in a Kripke semantics (with an 'intuitionistic logic'-like conditional).

Fourth, the Routley Star is there to deal with negation. The idea is to separate the falsemaking of φ (i.e.,$s \vDash \neg\varphi$) from the not-truthmaking of φ (i.e., $s \nvDash \varphi$): For falsemaking we need to consider the possible distinct twin world $s*$, which may be independent of the not-truthmaking at the present world s. This is similar to the semantics for FDE in terms of separate truthmaking and falsemaking. And to some extend this also works here, but the Routley Star is more general. We could also ignore the Routley Star and work in the positive fragment of the language—i.e., where we don't have the negation symbol. One then gets the relevant logic B^+ which is independently interesting (since it is simpler but still has the essential features of a relevance logic).

Exercise 8.2. Construct Routley–Meyer models invalidating three irrelevancies $p \to (q \to q)$ and $p \to (q \vee \neg q)$ and $(p \wedge \neg p) \to q$.

Exercise 8.3. Show that the relevance logic has the *variable sharing property*: If $\vDash_B \varphi \to \psi$, then there is some atomic sentence p that occurs both in φ and in ψ.

There also is a rich correspondence theory: Restall (1993) collects many properties of frames and relates them to validities.

8.3 Assessment

Three points to be discussed:

to be written in full, for now see Berto and Jago (2019, ch. 6).

- Variable sharing property as a syntactic test for relevance (necessary but not claimed to be sufficient)
- Interpretation of ternary relation in terms of information and the 'just pure but not applied semantics' criticism
- A peek into substructural logics: Restall (2000).

8.4 Exercises

Exercise 8.a (Problem)**.** Do exercise 8.3.

8.5 Notes

Berto and Jago (2019, ch. 6), Restall (1993), and Priest (2008, ch. 10).

9 Paradoxes and theories of truth

This is the last chapter, and we come full circle: getting back to the paradoxes. We started out with them and now have many more tools to tackle them. We first unveil the unifying structure behind the paradoxes. Then we focus on the liar paradox: how a theory of truth that avoids it can look like (e.g., Kripke's theory of truth).

Key concepts
- Inclosure scheme
- Generalized Cantor's theorem
- Fixed point theorem
- Tarski's hierarchy of languages
- Typed vs. type-free theories of truth
- Kripke's theory of truth

9.1 The unifying structure of self-referential paradoxes

Looking back at the paradoxes mentioned at the beginning (chapter 1), it seems that many of them have a common theme: self-reference. In some way or other, this seems to be behind, e.g., the liar paradox, the revenge paradox, the card paradox, Curry's paradox—and you might add many more yourself. However, it is notoriously difficult to precisely formulate this unifying structure of the self-referential paradoxes (see Priest 1994). In this section, we consider two fruitful attempts:

1. Inclosure scheme: The elaboration of Russell's idea by Priest (1994, 2010b).

2. Generalized Cantor's theorem: The elaboration of Lawvere's theorem by Yanofsky (2003).

We show that the second approach is more (or at least equally) general: the first is an instance of the second (example 9.8 below).

These papers at least don't cross-reference each other: so this might be something new.

One can also discuss whether other paradoxes that aren't (obviously) self-referential also have this unifying structure. The one important such paradox that we've discussed here is the sorites paradox. It has a treatment on the first approach (Priest 2010b). And with this connection, it thus also gets one on the second approach—which was missing so far.

9.1.1 Inclosure scheme

Early on, Russell had the following idea of identifying the unifying structure of self-referential paradoxes:

> Given a property φ and a function δ, such that, if φ belongs to all members of X, $\delta(X)$ always exists, has the property φ, and is not a member of X; then the supposition that there is a class Ω of all terms having property φ and that $\delta(\Omega)$ exists leads to the conclusion that $\delta(\Omega)$ both has and has not the property φ (B. Russell 1905b, p. 35, notation adjusted to accord with the recent literature).

Priest (1994, 2010b) worked out this idea into the *inclosure schema*: Think of δ as a construction that builds, given a set X of φ-objects, a new φ-object $\delta(X)$—where 'new' means $\delta(X) \notin X$. We get in trouble at the limit of this construction: whenever we think we reached a fixed point Ω where we constructed all the possible new φ-objects, we get the contradiction that $f(\Omega)$ is a new object we haven't yet considered. More formally:

Theorem 9.1 (Inclosure schema)**.** *Assume φ and θ are unary predicates and δ a partial function. Then the following are jointly inconsistent:*

1. *There is a set Ω such that $\Omega = \{x : \varphi(x)\}$ and $\theta(\Omega)$*
2. *If $X \subseteq \Omega$ and $\theta(X)$, then $\delta(X)$ exists and (a) $\delta(X) \notin X$ (b) $\delta(X) \in \Omega$.*

Proof. Consider $X := \Omega$. Then $\delta(\Omega) \notin (\Omega)$ and $\delta(X) \in \Omega$. □

If this is supposed to exhibit the unifying structure of paradoxes, we need to demonstrate how the well-known paradoxes are instances of this inclosure scheme. We do this here for Russell's paradox, the liar paradox, and the sorites paradox.

Or try yourself. It's fun!

Example 9.2 (Russell's paradox)**.** Russell's paradox shows that the set of all sets that don't contain themselves cannot exist. So take $\varphi(x)$ as '$x \notin x$', take a trivial $\theta(x)$ (e.g., '$x = x$'), and let δ be the identity function. Assume for contradiction that the class/collection $\{x : x \notin x\}$ of sets that don't contain themselves is a set Ω—so (1) holds. Also (2) holds: if $X \subseteq \Omega$, then $X \in X$ implies $X \in \Omega$, i.e., $X \notin X$; so $X \notin X$, and hence also $X \in \Omega$. Thus, theorem 9.1 indeed yields contradiction. ⌟

See Priest (1994, p. 27).

Example 9.3 (The liar paradox)**.** The liar sentence shows that the sentence 'This sentence is false' cannot have a classical truth-value: it is true iff it is

See Priest (1994, p. 30).

false. So take $\varphi(x)$ as 'x is true', take $\theta(x)$ as 'x is definable', and let δ be the partial function, obtained by some suitable diagonalization technique, that maps a definable set X to the sentence 'This sentence is not in X'. (The assumption that X is definable ensures that we can form this sentence.) Let $\Omega = \{x : \varphi(x)\}$ be the set of true sentences. Qua syntactic objects, there are no'size-issues', so this is indeed a set, so (1) holds. Also (2) holds: if $X \subseteq \Omega$ is definable, then $\delta(X)$ is defined, and if we had $\delta(X) \in X$, then

$$\delta(X) = \text{'This sentence is not in } X\text{' is in } X \subseteq \Omega\text{, i.e., is true}$$

so what the sentence $\delta(X)$ says is the case, i.e., $\delta(X) \notin X$, contradicting the assumption. So $\delta(X) \notin X$. Hence what the sentence $\delta(X)$ is the case, so $\delta(X)$ is true, i.e., $\delta(X) \in \Omega$. So theorem 9.1 indeed yields contradiction. ⌟

Example 9.4 (Sorites paradox)**.** In the sorites paradox, we consider a sequence of objects $a_1, \ldots, a_n$ (e.g., collections of grains of sand) and a vague predicate φ (e.g., 'is a heap'). These objects are such that a_1 (clearly) is φ and a_n (clearly) is not φ. But adjacent objects a_i and a_{i+1} are so similar that if one is φ, so is the other. That's known as the principle of tolerance. The paradox is that repeated application of tolerance yields that actually a_n is φ after all. This can be seen as an instance of the inclosure schema as follows. Write $A = \{a_1, \ldots, a_n\}$. Let θ be a trivially true property, and let δ be the function that is defined on proper subsets X of A and maps them to the first a_i in the sequence $a_1, \ldots, a_n$ that is not in X. Let $\Omega = \{a \in A : a \text{ is } \varphi\}$ be the set of objects that are φ, so (1) holds. Also (2) holds: If $X \subseteq \Omega$, then X is a proper subset of A (since $a_n \notin \Omega$), so $\delta(X)$ is defined and, by definition, not in X. Moreover, $\delta(X) \in \Omega$, because if $X = \emptyset$, then $\delta(X) = a_1 \in \Omega$, and if $X \neq \emptyset$, then $\delta(X)$ comes immediately after something in $X \subseteq \Omega$, so, by tolerance, $\delta(X) \in \Omega$. Thus, the 'diagonalization' construction δ takes us just outside a set of φ-objects, and tolerance keeps us within the set Ω of φ-objects; but, at the limit Ω of φ-objects, a contradiction hence must arise. ⌟

See Priest (2010a, 70 f.).

9.1.2 Generalized Cantor's theorem

The second, seemingly independent approach to finding a unified structure behind self-reference was provided by Lawvere (1969), introducing it as follows:

> The similarity between the famous arguments of Cantor, Russell, Gödel and Tarski is well-known, and suggests that these arguments should all be special cases of a single theorem about a suitable kind of abstract structure (Lawvere 1969, p. 134).

Lawvere went on to precisely provide such a theorem. It is formulated in the language of category theory (specifically for Cartesian closed categories). For accessibility, we follow here the elaboration of Yanofsky (2003) and state the result only in the language of sets and functions. (We mention below what difference this makes.)

If you want to read up on the category-theoretic result, with applications in theoretical computer science, I recommend Yanofsky (2022).

We call the result *generalized Cantor's theorem*. (Yanofsky also does so, but eventually goes for just 'Cantor's theorem' following a book by Lawvere and Schanuel.) So, as a motivation, let's start with the original Cantor's theorem.

- *Cantor's theorem*: There is no surjective function $\mathbb{N} \to 2^{\mathbb{N}}$, where $2 = \{0, 1\}$ and $2^{\mathbb{N}}$ is the set of functions from $\mathbb{N}$ to 2.

$2^{\mathbb{N}}$ can be identified with the powerset $\mathcal{P}(\mathbb{N})$ of $\mathbb{N}$.

- *Proof*: Assume for contradiction that there is a surjection $F : \mathbb{N} \to 2^{\mathbb{N}}$. Let's rewrite it as $f : \mathbb{N} \times \mathbb{N} \to 2$ with $f(n, m) := F(n)(m)$. Define $g : \mathbb{N} \to 2$ by $g(n) = \neg f(n, n)$, where $\neg : 2 \to 2$ maps 0 to 1 and 1 to 0. Since F is surjective, there is $n \in \mathbb{N}$ with $g = F(n)$, i.e., $g(-) = f(n, -)$. But then $f(n, n) = g(n) = \neg f(n, n)$, contradicts $\neg$ not having fixed points (i.e., $t \in 2$ with $\neg(t) = t$).

Visualize this proof and see why it is called a 'diagonal argument'. Start with

f	*0*	*1*	*2*	...
0				
1				
2				
⋮				

We now generalize this to any set of objects X (instead of $\mathbb{N}$) and any set of truth-values T (instead of 2). The conclusion remains: The set X cannot exhaustively talk about the possible properties its elements can have; i.e., if we aim to represent properties (i.e., functions $X \to T$) by objects (i.e., elements of X), we will always miss out on some.

Theorem 9.5 (Generalized Cantor's theorem)**.** *Let T be a set and $\alpha : T \to T$ a function* without fixed points *(i.e., for all $t \in T$, $\alpha(t) \neq t$). Let X be a set and $\Delta : X \to X \times X$ the* diagonal function *(i.e., $\Delta(x) = (x, x)$). Let $f : X \times X \to T$ be a function. Define the function $g : X \to T$, in the following diagram, as $\alpha \circ f \circ \Delta$.*

$$\begin{array}{ccc} X \times X & \xrightarrow{f} & T \\ {\scriptstyle\Delta}\uparrow & & \downarrow{\scriptstyle\alpha} \\ X & \xrightarrow[g]{} & T \end{array}$$

Then g is not representable *by f, i.e., for all $y \in X$, $g(-) \neq f(-, y)$.*

Proof. If there were such $y \in X$, then, by representability and definition,

$$f(y, y) = g(y) = \alpha\big(f(y, y)\big),$$

contradicting α not having fixed points. □

Now, again, if this is to identify the structure of paradoxes, we need to provide instances. For comparability, let's consider the liar paradox. For many more, see Yanofsky (2003).

Example 9.6 (The liar paradox)**.** Let $T = 2 = \{0, 1\}$ and $\neg : 2 \to 2$ classical negation (which has no fixed points). Let X be the set of declarative English sentences. Define $f : X \times X \to T$ by

$$f(x, y) := \begin{cases} 0 & \text{if sentence } y \text{ says that sentence } x \text{ is false} \\ 1 & \text{otherwise.} \end{cases}$$

By the generalized Cantor's theorem, we get the function $g : X \to T$ where

$$g(x) = 1 \text{ iff } f(x, x) = 0 \text{ iff } x \text{ says that it is false.}$$

In this sense, g is the characteristic function of liar sentences. The fact that g is not representable means that we cannot express the truth of (precisely the) liar sentences: there is no sentence $y \in X$ such that y says that x is true iff x is a liar sentence. ⌟

Exercise 9.7. Go through other paradoxes with the generalized Cantor's theorem (e.g., those analyzed using the inclosure scheme before, or yet other ones).

As mentioned, we want to show that the 'generalized Cantor' approach is more (or at least equally) general than the 'inclosure' approach: i.e., that the latter is an instance of the former.

In fact, also the inclosure approach can instantiate the generalized Cantor approach: exercise 9.b. So they are equally general.

Example 9.8 (Inclosure schema)**.** Assume φ and θ are unary predicates and δ a partial function. Let's assume condition (1) of the inclosure schema: Ω is a set such that $\Omega = \{x : \varphi(x)\}$ and $\theta(\Omega)$. Using the generalized Cantor's theorem, we derive a contradiction from condition (2) of the inclosure schema:

$(*)$ If $U \subseteq \Omega$ and $\theta(U)$, then $\delta(U)$ exists and (a) $\delta(U) \notin U$ (b) $\delta(U) \in \Omega$.

Let $T := 2$ and $\alpha := \neg : 2 \to 2$ mapping 1 to 0 and 0 to 1 (which has no fixed points). Let $X := \{U \subseteq \Omega : \theta(U)\}$. Define $f : X \times X \to 2$ by $f(U, V) = 1$ iff $\delta(U) \in V$. (Note that, if $U \in X$, then, by $(*)$, $\delta(U)$ is defined.) Now, the generalized Cantor's theorem applies yielding that $g := \neg \circ f \circ \Delta$ is not representable. In particular, since Ω is in X by assumption, we have $g(-) \neq f(-, \Omega)$. So there is $U \in X$ with $g(U) \neq f(U, \Omega)$. Note that $g(U) = 1$ iff $f(U, U) = 0$ iff $\delta(U) \notin U$. The latter holds by $(*)$, so we must have $f(U, \Omega) = 0$, so $\delta(U) \notin \Omega$, contradicting $(*)$. ⌟

We end this section with some comments to put the generalized Cantor's theorem into perspective.

First, also the present set-theoretic version can be stated more generally: We can replace the second X in $X \times X$ by some set Y, and replace Δ by the function $(\mathrm{id}, \beta) : X \to X \times Y$ with $\beta : X \to Y$ surjective; so f now is a function from $X \times Y$ to T. Thus, the parameters y in $f(-, y)$ aiming to represent g can be of a different kind than the elements of X, as long as there aren't more of them than elements in X.

Second, the contrapositive of the generalized Cantor's theorem is equally important, because it is a fixed point theorem. It is called the *diagonal theorem*. It says: If $f : X \times X \to T$ is a function such that all functions $g : X \to T$ are representable by f (i.e., there is $y \in X$ such that $g(-) = f(-, y)$), then all functions $\alpha : T \to T$ have a fixed point.

Third, what's missing in the set-theoretic setting as compared to the category-theoretic one is this: If we work in another category than the category of sets, then X and T can be sets with additional structure on them (e.g., partial orders, Boolean algebras, or topological spaces) and the functions Δ, f, and α are required to preserve this structure (i.e., are monotone, BA-homomorphisms, or continuous). Then it becomes harder for such functions to be free of fixed points: On a set T with two or more elements, a function without fixed points always exists, but such a function may fail to preserve the additional structure on T. Also, if we require functions from X to T to preserve structure, there are fewer such functions, so representability becomes easier. Thus, the diagonal theorem really gets its power in this category-theoretic setting.

Fourth, many famous theorems in logic and theoretical computer science can be proven using the generalized Cantor's theorem (and the diagonal theorem). This includes Gödel's incompleteness theorems, the Halting problem, Tarski's undefinability of truth, etc. We cannot go into this here for reasons of time, but the story is told by Yanofsky (2003).

Fifth, for a generalization beyond the category-theoretic result of Lawvere, to cover the Brandenburger–Keisler 'paradox' in epistemic game theory, see Abramsky and Zvesper (2012).

9.2 Theories of truth

Now that we've seen the unifying structure of paradoxes, let's look in detail at one of the most famous and ancient ones: the liar paradox. (The other famous ancient one, the sorites paradox, we already covered in

detail.) It calls for a theory of truth—i.e., an explanation of our concept of truth—which avoids contradiction. In this section, we describe some theories of truth. This includes Kripke's theory of truth. We motivate how it arises naturally from the generalized Cantor's theorem seen as a fixed point theorem.

9.2.1 Motivation

We've already seen two theories of truth. First, the correspondence theory: a sentence is true if it corresponds to a fact in the world. This comes natural in the context of classical logic, since it suggests bivalence: either a sentence does correspond to a fact or it doesn't. And, second, the coherence theory: roughly, a sentence is true if it is coherent with—or derivable from—the accepted sentences. This comes natural in the context of intuitionistic logic, since it doesn't want to presuppose the existence of a real world (especially in the context of math).

However, any theory of truth is challenged by the liar paradox, because it shows that only very few basic assumptions about how truth works (regardless of what it is) already lead to contradiction. So many theories of truth focus on providing a (formal) solution to the paradox: i.e., a notion of truth that is consistent and still has as much of the desired properties as possible. Although there are many (this really is a vast field), there isn't (yet) a universally accepted solution: each has some benefits but also disadvantages. Here we focus on a classic one: Kripke's theory of truth, but we mention several others along the way.

In fact, arguably the most classic theory of truth is the Tarskian way of defining truth (for formal languages). We sharply distinguish the object-language from the meta-language: In the object-language, we formulate our sentences, and in the meta-language we describe when they are true—namely, when the model in which we're interpreting them has the features that the sentence claims it to have (so that is a correspondence theory). However, the downside of this approach is that, in the object language, we *cannot* talk about truth. That is only something we can do in the metalanguage.

Indeed, Tarski's undefinability theorem shows that, in general, the meta-language really has to be stronger than the object-language. That is, even in expressive languages—e.g., the one to talk about arithmetic—we cannot cleverly define a truth-predicate which holds of (the numerical code of) a sentence iff the sentence is true in the model. This also can be seen as an instance of the generalized Cantor's theorem (Yanofsky 2003, p. 380).

But in natural language, we *can* talk about truth in the object-language. We say things like 'what they claim is true'. And also the liar sentence 'this sentence is false' is grammatical and, at least on first sight, also meaningful. One might think such talk is not really needed, since 'it is true that φ' just means 'φ'. But the truth-predicate still adds expressive power: now we can say things like 'all sentences of Peano arithmetic are true' even though there are infinitely many of them which we hence could not just list individually.

So it would be more satisfying to also have a theory of truth for languages that contain a truth-predicate. In fact, Tarski also considered how to do this. Facing the dilemma between having a language without its own truth-predicate and giving up classical logic, he opted for the first. But we don't need to stop at the meta-language (describing truth of the object-language). Truth in that meta-language can be described in a yet further meta-meta-language. Thus, we get a *hierarchy of languages* where each language can still talk about truth of sentences of the levels below. So the liar still cannot arise, since it denies the truth of a sentence at the current level.

More generally, this kind of approach is known as a *typed* theory of truth: the truth-predicate only applies, roughly, to sentences not containing the truth-predicate. While *type-free* theories of truth also allow such applications and hence also are called self-referential theories of truth.

However, this approach still cannot capture intuitive usage: Alice might say

> All Bob said today is false.

This places Alice's utterance to level n which is one higher in the hierarchy than the highest of Bob's utterances from today. But assume, unbeknownst to Alice, Bob said

> All Alice said today is true.

But then Bob's sentence must be on level $n + 1$, contradicting that all of Bob's utterances are of level $< n$. Thus, this fairly intuitive usage of the truth-predicate wouldn't be allowed.

So we'll now look at a type-free theory of truth. It chooses the other horn of the dilemma: considering a language with a truth-predicate but giving up some laws of classical logic (going three-valued).

9.2.2 Formal logic: Kripke's theory of truth

The generalized Cantor's theorem already suggested a close connection between fixed points and the liar paradox (example 9.6): Roughly, expressability of the liar sentence implied a fixed point for the negation operator. Since negation on two truth-values cannot have fixed points, it is suggestive to consider a third truth-value. This is what we'll see in Kripke's theory. Let's sketch it step by step.

We follow Restall (2022).

The language and the liar sentence First, we need to fix the language for which we want to develop a theory of truth. That's not a trivial matter. To focus on the main ideas, we won't be completely precise here. (But we give some pointers to a more precise treatment below.) The idea is this:

As 'base language' we use our usual language $\mathcal{L}_{\text{prop}}$ (using the connectives $\neg, \vee, \wedge, \bot, \top, \rightarrow, \leftrightarrow$). To get our desired language $\mathcal{L}^{\mathsf{T}}$, we add:

1. a *truth-predicate* T: So $\mathsf{T}x$ intuitively says that object x is true. More precisely, the symbol 'T' is applied to names (or singular terms) that denote the objects which are true or untrue. Typically, these objects x are sentences, but could also be other truth-bearers.

2. a *quotation* operator $\ulcorner\cdot\urcorner$: if φ is a sentence, then $\ulcorner\varphi\urcorner$ is a name (or singular term) for this sentence. So it becomes something to which the truth-predicate can apply: The claim 'Sentence φ is true' hence can be formalized as '$\mathsf{T}\ulcorner\varphi\urcorner$'.

3. And we assume we have a *liar* sentence λ. That is, λ is the sentence that says of itself that it is not true. In symbols: $\lambda = \ulcorner\neg\mathsf{T}\lambda\urcorner$, i.e., the two singular terms λ and $\ulcorner\neg\mathsf{T}\lambda\urcorner$ denote the same object, namely the liar sentence.

Reflect on this formalization of the intuitive description of the liar sentence.

Some comments to motivate this choice.

First, why the complication of using a predicate and quotation? Can't we simply take T as a unary connective? Sure, the natural language expression '… is true' makes T look like a predicate (assigning a property to objects). But couldn't we also say 'it is true that …' making T look like a sentential operator (taking a sentence and producing a new one)? However, then we cannot express the following that we would like to express with a truth-predicate:

See Leitgeb (2007, p. 277).

- The last sentence spoken by the Queen is true. (Even though we might not know which sentence this was.)

- All sentences of Peano arithmetic are true. (Even though we cannot list them all.)
- For all x and y, if y is the negation of x, then Tx if and only if not Ty.

See Leitgeb (2007, p. 277).

Second, what are the objects x to which we apply the truth-predicate? In principle, they could be any truth-bearer: sentences, propositions, utterances, etc. But, of those, sentences are philosophically best understood: they are syntactic objects. So this is the common choice of modern theories of truth. But we then should also have in our background language our theory of syntax describing how sentences are formed. To achieve this, the literature usually uses the trick of 'Gödelization': encoding sentences effectively by natural numbers. Thus, the theory of arithmetic can be used to describe syntax. And this theory is logically well understood and goes by the name of Peano arithmetic.

Third, this then provides a way to define $\mathcal{L}^T$ formally: One starts, as base language, with the first-order language of arithmetic containing, in addition to the logical symbols, the non-logical symbols $0, S, +, \times$. We add to this a unary predicate T. As axioms governing these symbols we uses Peano arithmetic PA. Formulas φ are coded by numbers/numerals $\ulcorner\varphi\urcorner$, so it is grammatical to write $T\ulcorner\varphi\urcorner$. Then one shows *Gödel's Diagonalization Lemma*. (This can also be done using the generalized Cantor's theorem (Yanofsky 2003, 378 f.).) This says, roughly, that for any predicate $F(x)$, there is a sentence φ such that φ is equivalent to $F(\ulcorner\varphi\urcorner)$. So we'd take $F(x) = \neg Tx$ and get $\lambda := \varphi$ with λ being equivalent to $\neg T\ulcorner\lambda\urcorner$. So, in a sense, λ says of itself that it is not true. Kripke (1975) highlights the upshot:

> "In this way, Gödel put the issue of the legitimacy of self-referential sentences beyond doubt; he showed that they are as incontestably legitimate as arithmetic itself" (Kripke 1975, p. 692).

Note a subtlety, though. For our liar sentence we assumed $\lambda = \ulcorner\neg T\lambda\urcorner$, not the equivalence of λ and $\neg T(\ulcorner\lambda\urcorner)$. For this we need the strong diagonal lemma (e.g. Heck Jr 2007, p. 7).

Models The task of finding a consistent theory of truth means, at least to a good approximation, finding a consistent interpretation of our language $\mathcal{L}^T$. So we want to find a *model* of $\mathcal{L}^T$. To deserve the name, such a model m should assign the $\mathcal{L}^T$-sentences truth-values in a way that respects

1. the meaning of the connectives

2. the defining feature of the truth-predicate: the T-*schema* saying that $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$ is equivalent to φ.

Before showing how Kripke did this using three truth-values, let's recall why it doesn't work with the two classical truth-values: Assume $\mathfrak{m} : \mathcal{L}^{\mathsf{T}} \to \{0, 1\}$ would be a model. To respect $\neg$, we should have $\mathfrak{m}(\neg\varphi) = \neg\mathfrak{m}(\varphi)$ (where the second '$\neg$' is the function mapping 1 to 0 and 0 to 1). To respect the T-scheme, we should have $\mathfrak{m}(\varphi) = \mathfrak{m}(\mathsf{T}^{\ulcorner}\varphi^{\urcorner})$. But then we have for the liar sentence (glossing over differences between formulas and singular terms denoting them):

$$\mathfrak{m}(\lambda) = \mathfrak{m}(\neg\mathsf{T}\lambda) = \neg\mathfrak{m}(\mathsf{T}\lambda) = \neg\mathfrak{m}(\lambda), \tag{9.1}$$

which is a contradiction since $\neg$ doesn't have a fixed point on $\{0, 1\}$.

As already noted, this suggests adding a third truth-value $\mathfrak{i}$ for 'undefined' and giving λ this truth-value. Then equation (9.1) is no contradiction anymore, because for the standard strong Kleene meaning of $\neg$ we have the fixed point $\neg(\mathfrak{i}) = \mathfrak{i}$.

But then we might ask: why can't we simply inductively define a model? On atomic sentences, let $\mathfrak{m}$ take any value in $\{0, 1, \mathfrak{i}\}$; for the propositional connectives use, say, the strong Kleene truth-tables; and for formulas of the form $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$ use the T-scheme: $\mathfrak{m}(\mathsf{T}^{\ulcorner}\varphi^{\urcorner}) = \mathfrak{m}(\varphi)$.

However, the problem is this. We're now attempting a recursion on names for formulas, but these are not well-ordered by their complexity as the set of formulas is. Concretely, the liar sentence breaks this well-order: We define the value of $\mathsf{T}\lambda$ by recursing to the value of λ. But we need to make sure that it has the same value as $\neg\mathsf{T}\lambda$. Though, we cannot consider this yet, since it is of higher complexity.

Kripke's trick is to move away from the *local* perspective of considering a single model: trying to determining the value of a sentence by recursion to less complex sentences. Rather, take a *global* perspective of considering all models: how we can improve a given model and thus produce a new model that is closer to satisfying the T-schema.

Kripke's theory of truth We construct a three-valued model $\mathfrak{m}_*$ of $\mathcal{L}^{\mathsf{T}}$ as follows using three key ideas. (Again, the presentation is not fully precise, but rather focuses on intuition.)

Idea 1: Update

We start with a model $\mathfrak{m}_0$ in which the T-free sentences get any value we like (corresponding to 'the actual world'), but all sentences containing T get the value $\mathfrak{i}$. So $\mathfrak{m}_0$ doesn't yet satisfy the T-schema because, for a true

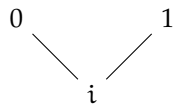

Figure 9.1: The order $\leqslant_I$ of $\{0, i, 1\}$ by information.

atomic sentence p, the sentence $T^\ulcorner p^\urcorner$ is still undetermined: $m_0(p) = 1 \neq i = m_0(T^\ulcorner p^\urcorner)$. We want to fix this.

So, given a model m, we build a new model $\rho(m)$ as follows: For T-free sentences, $\rho(m)$ gives the same value as m; but sentences of the form $T^\ulcorner \varphi^\urcorner$ get the value $m(\varphi)$. This now avoids the recursion issue from before: qua model, m has a value for every sentence already, so we can refer to $m(\varphi)$.

The new model $m_1 := \rho(m_0)$ is better: We have $m_1(T^\ulcorner p^\urcorner) = m_0(p) = m_1(p)$. But we're not there yet: for the sentence $\varphi := T^\ulcorner p^\urcorner$, we have $m_1(T^\ulcorner \varphi^\urcorner) = m_0(\varphi) = i \neq 1 = m_1(\varphi)$. So we want to update again, but to be sure that this helps, we need the next key idea.

Idea 2: Informativeness

The crucial insight is that this update process ρ increases (and preserves) *informativeness*: defined formally as follows. We can order the three truth-values $0, 1, i$ by information using the order $\leqslant_I$ visualized in figure 9.1. The value i carries no information, but the values 1 and 0 each convey a maximal amount of consistent information. And this naturally extends to models: model n is at least as informative as model m (in symbols $m \leqslant n$) iff, for all sentences φ, the n-value of φ is at least as informative as the m-value. Formally:

We discussed this order already in figure 3.1 of chapter 3.

$$m \leqslant n :\Leftrightarrow \forall \varphi \in \mathcal{L}^T : m(\varphi) \leqslant_I n(\varphi).$$

Thus, the mentioned insight phrased formally is that ρ is inflationary: $m \leqslant \rho(m)$ (increasing informativeness). And it also is monotone: if $m \leqslant n$, then $\rho(m) \leqslant \rho(n)$ (preserving informativeness). This makes intuitive sense and, in a precise formal setting, is proved by induction on the formulas φ. It works not just for strong Kleene, but also for other logics with appropriate truth-tables.

Idea 3: Fixed point

Intuitively, we now apply the update over and over again to our starting model m_0. Thus, we get models of increasing informativeness:

$$m_0 \leqslant \rho(m_0) =: m_1 \leqslant \rho(m_1) =: m_2 \leqslant \rho(m_2) =: m_3 \leqslant \dots.$$

And we update until there is nothing more to update, i.e., until we reached

a fixed point: a model $\mathfrak{m}_*$ such that $\rho(\mathfrak{m}_*) = \mathfrak{m}_*$. Mathematically, the existence of such a fixed point is guaranteed by fixed point theorems. There are many such theorems and they play an important role in mathematics. In an order-theoretic context (like the present one), one makes crucial use of the fact that the function ρ for which one wants to find a fixed point is inflationary or monotone. We detail such a fixed point theorem in the appendix (theorem 10.1). But what really matters now is that such a fixed point model $\mathfrak{m}_*$ provides a model for $\mathcal{L}^{\mathsf{T}}$ that we were looking for: (1) qua model, $\mathfrak{m}_*$ respects the meaning of the connectives, and (2) qua fixed point, $\mathfrak{m}_*$ satisfies the T-schema:

Which I think deserves more attention.

$$\mathfrak{m}_*(\mathsf{T}^\ulcorner\varphi^\urcorner) = \rho(\mathfrak{m}_*)(\mathsf{T}^\ulcorner\varphi^\urcorner) = \mathfrak{m}_*(\varphi).$$

On a more philosophical take, we can summarize the main idea of Kripke's theory as distinguishing the *grounded* applications of the truth-predicate from the problematic *ungrounded* ones: A sentence like 'It is true that snow is white' is grounded because the truth-predicate applies to a fact—which always is grounded. The liar sentence 'This sentence is not true' is not grounded because the truth-predicate applies to a sentence (namely the whole sentence again) which is not yet grounded. So we can see the three truth-values as: grounded-and-true (1), grounded-and-false (0), and ungrounded (i). And we can see the updating process as the process of grounding more and more sentences. A fixed point of this process then determines which sentences eventually get grounded and which remain ungrounded.

Outro

9.2.3 Assessment

Here are two main criticisms of Kripke's theory. First, the liar sentence $\lambda = {}^\ulcorner\neg\mathsf{T}\lambda^\urcorner$ is saying of itself that it is not true. But, in a three-valued setting, this is not saying that it is false. That is, we do not also have a falsity-predicate available. And if we had, we'd get paradox again due to the revenge paradox: 'this sentence is either false or neither-true-nor-false'.

Second, Kripke's theory doesn't have a conditional which satisfies the deduction theorem, so the conditional cannot reflect meta-language reasoning. We cannot, for example, change it to the Łukasiewicz conditional, because that wouldn't be monotone anymore. (Field has introduced a new conditional with the deduction theorem, but that is quite tricky.)

There are many more theories of truth: For example, in addition to paracomplete theories like Kripke's (basing it on strong Kleene logic), there

are also paraconsistent logics (basing it on the logic of paradox). There also are supervaluationist approaches ($\mathsf{T}(\varphi)$ is true iff φ is supertrue). It also has been investigated if one of the directions of the transparency of truth ($\varphi \dashv\models \mathsf{T}(\varphi)$) can be given up. And structural approaches to truth explore if giving up some structural rules (like weakening/monotonicity, reflexivity, etc. of the classical consequence relation $\models$) is a viable option to block the liar reasoning.

Finally, theories of truth like Kripke's try to provide a (semantic) way of defining the truth-predicate. Axiomatic theories of truth, on the other hand, take the truth-predicate as a primitive notion and then consider various axioms for it. Thus, insights about truth are obtained by exploring and comparing different axiomatizations.

9.3 Exercises

Exercise 9.a. We said that the sorites paradox was not treated with the generalized Cantor's theorem by Yanofsky (2003). Can you do it? You can come up with an application yourself, or you can use the treatment of the sorites paradox using the inclosure scheme (example 9.4) together with the translation into the generalized Cantor's theorem (example 9.8).

Exercise 9.b. Also the generalized Cantor's theorem can be seen as an instance of the inclosure scheme. Can you find how? Are you satisfied with this instantiation?

9.4 Notes

Priest (1994), Priest (2010b), Yanofsky (2003), Kripke (1975), van Rooij (n.d.), Gauker (2006), Halbach (2011), J. c. Beall (2016), Restall (2022).

10 Appendix: some set theory and order theory

Set theory See, e.g., Priest (2008, sec. 0.1). Some additional comments:

Has to be written properly.

- A function f from a set A (its domain) to a set B (its codomain) is written $f : A \to B$. To say that f maps an element $a \in A$ to $b \in B$ we write $f(a) = b$ or, if f is clear from context, $a \mapsto b$.
- If $f : A^n \to A$ is a function (where $A^n = A \times \ldots \times A$ is the n-time Cartesian product), we say f has *arity* n (i.e., it takes n arguments from A to produce another element from A). Similarly, a sentential connective c that takes as arguments n sentences to produce a new sentence also is said to have arity n. The first arities have special names: unary (= 1-ary), binary (= 2-ary), and ternary (= 3-ary). Sometimes it is convenient to take a 0-ary function or connective to be a constant (i.e., an element or symbol which is fixed throughout).

Order theory Let S be a set and $\leqslant \subseteq S \times S$ a binary relation. We say $(S, \leqslant)$ is a *preorder* if $\leqslant$ is reflexive ($\forall x \in S : x \leqslant x$) and transitive ($\forall x, y, z \in S :$ $x \leqslant y$ and $y \leqslant z \Rightarrow x \leqslant z$). If $\leqslant$ is also antisymmetric ($\forall x, y \in S : x \leqslant$ y and $y \leqslant x \Rightarrow x = y$), we call $(S, \leqslant)$ a partial order. One defines $x < y$ as $x \leqslant y$ and $x \neq y$.

Finite partial orders can be represented by *Hasse diagrams*: the elements of S become points in the diagram, and there is a line from x to y if $x < y$ and there is no other element between x and y (i.e., there is no z with $x < z < y$). If this holds, one also says that y *covers* x. For example, the order $(\mathcal{P}(\{2, 5\}), \subseteq)$ of the subsets of the two element set $\{2, 5\}$ ordered by inclusion is described by the Hasse diagram of figure 10.1. It often is very helpful to think of partial orders visually as Hasse diagrams. Since they only depict the covering relation, they are less cluttered, but from it one can recover the partial order as the reflexive and transitive closure of the covering relation (i.e., add $x \leqslant x$, and add $x \leqslant z$ whenever $x \leqslant y$ and $y \leqslant z$ is depicted).

There are various notions of bounds: Let $(S, \leqslant)$ be a preorder (which includes partial orders). If $A \subseteq S$ is a subset, an *upper bound* (resp., *lower bound*) of A is an element $x \in S$ (if it exists) such that, for all $a \in A$, we have $x \geqslant a$ (resp., $x \leqslant a$). Moreover, x is the *least upper bound* (resp., *greatest*

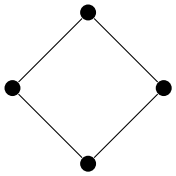

Figure 10.1: The Hasse diagram of $(\mathcal{P}(\{2,5\}), \subseteq)$.

lower bound) of A if x is an upper bound (resp., lower bound) of A and, for all upper bounds (resp., lower bounds) y of A, we have $x \leqslant y$ (resp., $x \geqslant y$). Note that, if least upper bounds exist, they are unique: i.e., if x and x' are least upper bounds of A, then $x = x'$—so we can speak of *the* least upper bound. Similarly for greatest lower bounds.

There are various notions of 'extremal' elements: Let $(S, \leqslant)$ be a preorder (which includes partial orders). An element $x \in S$ is the *greatest* (resp., *least*) element of S if, for all $s \in S$, $x \geqslant s$ (resp., $x \leqslant s$). Again, if existent, greatest and least elements are unique. An element $x \in S$ is a *maximal* (resp., *minimal*) element of S if, for all $s \in S$, if $x \leqslant s$ (resp., $s \leqslant x$), then $x = s$. So there is no other element strictly above (resp., below) x. Note that being maximal (resp., minimal) doesn't imply being greatest (resp., least): a partial order can have several maximal (resp., minimal) elements. (Think of some examples.)

Fixed point theorems in order theory There are many fixed point theorems in order theory (and in math in general). In order theory, they provide conditions when a function $f : S \to S$ on a partial order $(S, \leqslant)$ has a fixed point, i.e., a point $s \in S$ with $f(s) = s$. Probably the most famous one is the Knaster–Tarski theorem (requiring f to me monotone and S to be a complete lattice). A much less well-known strengthening (mentioned in chapter 9) is Pataraia's theorem. (We follow the nlab proof.)

The citation is Pataraia (1997), though I couldn't get a hold of the paper. Also see Martin (2013).

First some definitions: Let $(S, \leqslant)$ be a partial order. A subsets $A \subseteq S$ is *directed* if A is nonempty and for any two elements of A have an upper bound in A (i.e., $\forall a, b \in A \exists c \in A : a, b \leqslant c$). The partial order S is *directed-complete* if any directed subset has a least upper bound (aka directed join). Moreover, a function $f : S \to S$ on a partial order $(S, \leqslant)$ is *monotone* if, for all $a, b \in S$, if $a \leqslant b$, then $f(a) \leqslant f(b)$.

Useful for chapter 9: the set $[X \to S]$ of functions from a set X to a directed-complete partial order S is again directed-complete under the pointwise order. If X is the set of sentences and S the information order of the three truth-values, then $[X \to S]$ is the set of models.

Theorem 10.1 (Pataraia)**.** *Let $(S, \leqslant)$ be a directed-complete partial order with a least element $\bot$. Then any monotone function $f : S \to S$ has a fixed point.*

One can strengthen the theorem to f having a *least* fixed point, but here

we only need its existence.

Proof. Consider the smallest subset D of S that contains $\bot$, is closed under f, and closed under directed joins (take D to be the intersection of all such subsets). Then the restriction of f to D is a function $f : D \to D$. And, for all $x \in D$, we have $x \leqslant f(x)$: This is because the set $E := \{x \in S : x \leqslant f(x)\}$ also contains $\bot$, is closed under f (if $x \in E$, then $x \leqslant f(x)$, so, by monotonicitiy, $f(x) \leqslant f(f(x))$, so $f(x) \in E$), and closed under directed joins (if $A \subseteq E$ is directed, then $\bigvee A \leqslant f(\bigvee A)$, because, for $a \in A$, we have $a \leqslant f(a) \leqslant f(\bigvee A)$, so $f(\bigvee A)$ is an upper bound of A). So $D \subseteq E$ since D is the smallest such set. Hence, for all $x \in D$, we have $x \in E$, i.e., $x \leqslant f(x)$.

A monotone function $g : D \to D$ with $x \leqslant g(x)$ for all $x \in D$ is called *inflationary* (it makes things bigger-or-equal than they were). We want to show that there is a largest inflationary function t on D. Consider the collection I of inflationary functions $g : D \to D$. Order it elementwise: $g \leqslant h$ iff, for all $x \in D$, $g(x) \leqslant h(x)$. The least element then is the identity function $\mathsf{id} : D \to D$ (since $\mathsf{id}(x) = x \leqslant g(x)$). And I is directed-complete: if $A \subseteq I$ is directed, the join is computed component-wise: $(\bigvee A)(x) = \bigvee_{a \in A} a(x)$ (since D is directed-complete). So $(I, \leqslant)$ again is a directed-complete partial order. In fact, I itself actually is directed: it is nonempty since our f is in I, and for $g, h \in I$, the function composition $g \circ h$ is in I (since $x \leqslant h(x) \leqslant g(h(x))$ such that $g \leqslant g \circ h$ (since $g(x) \leqslant g(h(x))$ because $x \leqslant h(x)$) and $h \leqslant g \circ h$ (since $h(x) \leqslant g(h(x))$). So I has a greatest element, call it t (like top).

This finishes the proof: We have $f \circ t \leqslant t$ because $f \circ t \in I$ (as above) and t is the greatest element. And we have $t \leqslant f \circ t$ since f is inflationary (so $t(x) \leqslant f(t(x))$). Hence $f \circ t = t$. Now $s := t(\bot) \in D \subseteq S$ is a fixed point of f: $f(s) = f(t(\bot)) = t(\bot)$. □

This is an amazingly simple and—importantly—constructive proof. There are similar results like the Bourbaki–Witt theorem: every inflationary function f on a nonempty chain-complete partial order S has a fixed point. But their usual proofs require some heavy machinery like that there is no injective mapping of the ordinals into a set: then the sequence starting with some element $s_0 \in S$, setting $s_{\alpha+1} = f(s_\alpha)$, and, for limit-ordinals α, taking s_α as the join of the chain $\{s_\beta : \beta < \alpha\}$, must eventually be constant. If, additionally, the function preserves joins (which is known as Scott continuity), then we don't need this machinery because then the sequence stops at ω: $f(s_\omega) = f\big(\bigvee\{s_0, s_1, \ldots\}\big) = \bigvee\{f(s_0), f(s_1), \ldots\} = s_\omega$. (This is known as the Kleene fixed point theorem.)

Bibliography

Abramsky, S. and J. Zvesper (2012). "From Lawvere to Brandenburger-Keisler: Interactive Forms of Diagonalization and Self-reference." In: *Coalgebraic Methods in Computer Science*. Ed. by D. Pattinson and L. Schröder. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–19 (cit. on p. 117).

Beall, J. c. (2016). "Off-Topic: A New Interpretation of Weak-Kleene Logic." In: *The Australasian Journal of Logic* 13.6. DOI: 10.26686/ajl.v13i6.3976 (cit. on pp. 30, 125).

Beall, J., M. Glanzberg, and D. Ripley (2020). "Liar Paradox." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University (cit. on pp. 5 sq.).

Beall, J. and G. Restall (2005). *Logical Pluralism*. Oxford: Oxford University Press (cit. on p. 27).

Belnap, N. D. (2019a). "A Useful Four-Valued Logic." In: *New Essays on Belnap-Dunn Logic*. Ed. by H. Omori and H. Wansing. Synthese Library 418. Cham: Springer, pp. 35–53 (cit. on pp. 31, 37 sq.).

— (2019b). "How a Computer Should Think." In: *New Essays on Belnap-Dunn Logic*. Ed. by H. Omori and H. Wansing. Synthese Library 418. Cham: Springer, pp. 55–76 (cit. on pp. 31, 37).

Berto, F. and M. Jago (2019). *Impossible Worlds*. Oxford: Oxford University Press. URL: https://global.oup.com/academic/product/impossible-worlds-9780198812791?cc=gb&lang=en& (cit. on pp. 110 sq.).

Berto, F. and D. Nolan (2021). "Hyperintensionality." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cit. on pp. 74, 83).

Bezhanishvili, N. and D. de Jongh (2006). *Intuitionistic Logic*. Prepublication Series PP-2006-25. Available at https://eprints.illc.uva.nl/

id/eprint/200/1/PP-2006-25.text.pdf (accessed 5 Nov 2021). Amsterdam (cit. on pp. 59, 66, 70).

Bjerring, J. C. and W. Schwarz (2017). "Granularity problems." In: *The Philosophical Quarterly* 67.266, pp. 22–37. URL: http://www.umsu.de/papers/granularity.pdf (cit. on p. 73).

Blackburn, P., M. Rijke, and Y. Venema (2001). *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge: Cambridge University Press (cit. on p. 13).

Blok, W. J. and D. Pigozzi (1989). *Algebraizable logics*. Vol. 77. Memoirs 396. Providence, Rhode Island, USA: American Mathematical Society (cit. on p. 19).

Bochvar, D. A. (1938). "On a Three- Valued Logical Calculus and Its Application to the Analysis of the Paradoxes of the Classical Extended Functional Calculus." In: *Mathematicheskii Sbornik* 4(46).2, pp. 287–308 (cit. on pp. 30, 130).

— (1981). "On a three-valued logical calculus and its application to the analysis of the paradoxes of the classical extended functional calculus." In: *History and Philosophy of Logic* 2. This is the English translation by M. Bergmann of the original Bochvar 1938, pp. 87–112 (cit. on p. 30).

Burgess, J. P. (1981). "Quick Completeness Proofs for Some Logics of Conditionals." In: *Notre Dame Journal of Formal Logic* 22.1, pp. 76–84 (cit. on pp. 89, 94, 96).

— (2009). *Philosophical Logic*. Princeton Foundations of Contemporary Philosophy. Princeton: Princeton University Press (cit. on p. 4).

Burris, S. and H. P. Sankappanavar (1981). *A Course in Universal Algebra*. An updated online version from 2012 is available at https://www.math.uwaterloo.ca/~snburris/htdocs/UALG/univ-algebra2012.pdf. New York: Springer (cit. on p. 17).

Cobreros, P. et al. (2012). "Tolerant, classical, strict." In: *Journal of Philosophical Logic* 41.2, pp. 347–385 (cit. on p. 36).

Cobreros, P. et al. (2015). "Vagueness, truth and permissive consequence." In: *Unifying the philosophy of truth*. Ed. by D. Achouriotti, H. Galinon, and J. Martinez. Dordrecht: Springer, pp. 409–430 (cit. on pp. 36, 40, 46).

— (2020). "Inferences and Metainferences in ST." In: *Journal of Philosophical Logic* 1057-1077, p. 49 (cit. on p. 36).

Dalla Chiara, M. L. (1986). "Quantum logic." In: *Handbook of philosophical logic*. Ed. by D. Gabbay and F. Guenthner. Vol. III. Synthese library 166. Dordrecht: D. Reidel Publishing Company, pp. 427–469 (cit. on p. 81).

Davey, B. A. and H. A. Pristley (2002). *Introduction to Lattices and Order*. 2nd ed. Cambridge: Cambridge University Press (cit. on p. 25).

David, M. (2020). "The Correspondence Theory of Truth." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University (cit. on p. 52).

Dunn, J. M. (1966). "The Algebra of Intensional Logics." PhD thesis. University of Pittsburgh (cit. on p. 30).

— (2019). "Two, Three, Four, Infinity: The Path to the Four-valued Logic and Beyond." In: *New Essays on Belnap-Dunn Logic*. Ed. by H. Omori and H. Wansing. Vol. 418. Synthese Library. Cham: Springer, pp. 77–97 (cit. on p. 30).

Dunn, M. J. and G. M. Hardegree (2001). *Algebraic Methods in Philosophical Logic*. Oxford Logic Guides 41. Oxford: Oxford University Press (cit. on pp. 4, 26).

Fine, K. (1975). "Critical Notice." In: *Mind* LXXXIV.1, pp. 451–458. DOI: 10.1093/mind/LXXXIV.1.451. URL: https://doi.org/10.1093/mind/LXXXIV.1.451 (cit. on p. 95).

— (2014). "Truth-Maker Semantics for Intuitionistic Logic." In: *Journal of Philosophical Logic* 43.2-3, pp. 549–577 (cit. on pp. 80, 83).

— (2016). "Angellic Content." In: *Journal of Philosophical Logic* 45.2, pp. 199–226 (cit. on pp. 80, 83).

Fine, K. (2017). "Truthmaker Semantics." In: *A Companion to the Philosophy of Language*. Wiley-Blackwell. Chap. 22, pp. 556–577 (cit. on pp. 74, 80, 83, 95).

Fine, K. and M. Jago (2019). "Logic for Exact Entailment." In: *The Review of Symbolic Logic* 12.3, pp. 536–556 (cit. on pp. 74, 82 sq.).

Font, J. M. (2016). *Abstract Algebraic Logic: An Introductory Textbook*. Studies in Logic 60. London: College Publications (cit. on pp. 4, 19).

Gauker, C. (07/2006). "Kripke's Theory of Truth (for the strong Kleene scheme)." Course notes. URL: http://www.christophergauker.sbg.ac.at/documents/KripkeTruth.pdf (cit. on p. 125).

Gehrke, M. (2009). *Duality*. Inaugurele Rede. URL: http://hdl.handle.net/2066/83300 (cit. on p. 12).

Gödel, K. (1932). "Zum Intuitionistischen Aussagenkalkül." In: *Anzeiger der Akademie der Wissenschaften in Wien* 69. In the "Kurt Gödel Collected Work" vol. 1, p. 222–225, pp. 65–66 (cit. on p. 57).

Grätzer, G. (2008). *Universal Algebra*. 2nd ed. New York, NY: Springer (cit. on p. 17).

Hájek, P. (1998). *Metamathematics of Fuzzy Logic*. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-011-5300-3 (cit. on p. 69).

— (2001). "On very true." In: *Fuzzy Sets and Systems* 124.3, pp. 329–333. DOI: https://doi.org/10.1016/S0165-0114(01)00103-8 (cit. on p. 68).

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambrdige: Cambridge University Press (cit. on p. 125).

Heck Jr, R. G. (2007). "Self-reference and the Languages of Arithmetic." In: *Philosophia Mathematica* 15.1, pp. 1–29. DOI: 10.1093/philmat/nkl028 (cit. on p. 121).

Hodges, W. (1983). "Elementary predicate logic." In: *Handbook of philosophical logic*. Vol. 1. Dordrecht: D. Reidel Publishing Company, pp. 1–131 (cit. on p. 26).

Hornischer, L. (2017). "Hyperintensionality and synonymy: a logical, philosophical, and cognitive investigation." Available at https://www.illc.uva.nl/Research/Publications/Reports/MoL-2017-07.text.pdf. MA thesis. Amsterdam: Institute for Logic, Language and Computation (cit. on p. 83).

— (2020). "Logics of Synonymy." In: *Journal of Philosophical Logic* 49, pp. 767–805. DOI: https://doi.org/10.1007/s10992-019-09537-5 (cit. on p. 83).

Hyde, D. and D. Raffman (2018). "Sorites Paradox." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University (cit. on p. 7).

Incurvati, L. and J. J. Schlöder (2021). "Meta-inferences and Supervaluationism." In: *Journal of Philosophical Logic*. DOI: https://doi.org/10.1007/s10992-021-09618-4 (cit. on p. 44).

Jago, M. (2014). *The Impossible. An Essay on Hyperintensionality*. Oxford: Oxford University Press. URL: http://dx.doi.org/10.1093/acprof:oso/9780198709008.001.0001 (cit. on p. 25).

Jansana, R. (2022). "Algebraic Propositional Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University (cit. on p. 19).

Johnstone, P. T. (1982). *Stone Spaces*. Cambrdige studies in advanced mathematics 3. Cambridge: Cambridge University Press (cit. on p. 20).

Khoo, J. (2015). "On Indicative and Subjunctive Conditionals." In: *Philosophers' Imprint* 15.32 (cit. on p. 91).

Kleene, S. C. (1952). *Introduction to Metamathematics*. Amsterdam: North-Holland (cit. on p. 46).

Kraus, S., D. Lehmann, and M. Magidor (1990). "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics." In: *Artifical Intelligence* 44, pp. 167–207 (cit. on pp. 99, 103, 106).

Kripke, S. (1975). "Outline of a Theory of Truth." In: *The Journal of Philosophy* 72.19, pp. 690–716 (cit. on pp. 121, 125).

Lawvere, F. W. (1969). "Adjointness in foundations." In: *Dialectica* 23. For a reprint, see http://www.tac.mta.ca/tac/reprints/articles/16/tr16abs.html (reprints in Theory and Applications of Categories), pp. 281–296 (cit. on p. 114).

Leitgeb, H. (2005). "Hodges' Theorem Does not Account for Determinacy of Translation. A Reply to Werning." In: *Erkenntnis* 62.3, pp. 411–425. DOI: 10.1007/s10670-004-1992-2 (cit. on p. 106).

— (2007). "What Theories of Truth Should be Like (but Cannot be)." In: *Philosophy Compass* 2.2, pp. 276–290 (cit. on pp. 120 sq.).

Lewis, D. (1971). "Completeness and decidability of three logics of counterfactual conditionals." In: *Theoria* 37.1, pp. 74–85. DOI: https://doi.org/10.1111/j.1755-2567.1971.tb00061.x (cit. on p. 90).

Lewis, D. K. (1973). *Counterfactuals*. Page numbers refer to the reissued version of 2001. Oxford: Blackwell (cit. on pp. 89 sq.).

MacBride, F. (2021). "Truthmakers." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University (cit. on p. 74).

Martin, K. (2013). "Nothing Can Be Fixed." In: *Computation, Logic,Games, and Quantum Foundations: The Many Facets of Samson Abramsky*. Ed. by B. Coecke, L. Ong, and P. Panangaden. Berlin Heidelberg: Springer-Verlag, pp. 195–196 (cit. on p. 127).

Menzel, C. (2021). "Possible Worlds." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University (cit. on p. 12).

Moschovakis, J. (2021). "Intuitionistic Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University (cit. on p. 66).

Nolan, D. (2014). "Hyperintensional metaphysics." In: *Philosophical Studies* 171.1, pp. 149–160 (cit. on p. 72).

O'Connor, J. J. and E. F. Robertson (02/2005). "Philip Edward Bertrand Jourdain." In: *MacTutor History of Mathematics Archive*. Available at https://mathshistory.st-andrews.ac.uk/Biographies/

Jourdain/ (accessed 22 Oct 2021). University of St Andrews, Scotland (cit. on p. 6).

OpenLogicProject (2021). *The Open Logic Text*. Revision: 00873f4 (master) 2021-10-25. URL: https://builds.openlogicproject.org/open-logic-complete.pdf (cit. on pp. 4, 53).

Pagin, P. (2017). "Tolerance and higher-order vagueness." In: *Synthese* 194.10, pp. 3727–3760 (cit. on p. 68).

Pataraia, D. (11/1997). "A constructive proof of Tarski's fixed-point theorem for dcpo's." In: *65th Peripatetic Seminar on Sheaves and Logic*. Aarhus, Denmark (cit. on p. 127).

Priest, G. (2010a). "The logic of the catuskoti." In: *Comparative Philosophy* 1.2, pp. 24–54 (cit. on p. 114).

Priest, G. (1979). "The Logic of Paradox." In: *Journal of Philosophical Logic* 8, pp. 219–241 (cit. on p. 34).

— (1994). "The Structure of the Paradoxes of Self-Reference." In: *Mind* 103.409, pp. 25–34. URL: https://www.jstor.org/stable/2253956 (cit. on pp. 112 sq., 125).

— (2008). *An Introduction to Non-classical Logic. From If to Is*. 2nd ed. Cambridge: Cambridge University Press (cit. on pp. 4 sq., 26, 29, 31, 35, 37, 43, 45, 47, 49, 67, 69 sq., 90, 96, 107 sq., 111, 126).

— (2010b). "Inclosures, Vagueness, and Self-Reference." In: *Notre Dame Journal of Formal Logic* 51.1, pp. 69–84. DOI: 10.1215/00294527-2010-005 (cit. on pp. 112 sq., 125).

Restall, G. (1993). "Simplified semantics for relevant logics (and some of their rivals)." In: *Journal of Philosophical Logic* 22, pp. 481–511. DOI: https://doi.org/10.1007/BF01349561 (cit. on pp. 109 sqq.).

— (2000). *An Introduction to Substructural Logics*. London: Routledge (cit. on pp. 3 sq., 80, 111).

— (2022). *Proofs and Models in Philosophical Logic*. Elements in Philosophy and Logic. Cambridge: Cambridge University Press (cit. on pp. 4, 120, 125).

Russell, B. (1905a). "On Denoting." In: *Mind* 14.56, pp. 479–493 (cit. on pp. 5, 29).

— (1905b). "On Some Difficulties in the Theory of Transfinite Numbers and Order Types." In: *Proceedings of the London Mathematical Society* 4.14, pp. 29–53. DOI: 10.2307/2011035 (cit. on p. 113).

Russell, G. (2021). "Logical Pluralism." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cit. on p. 27).

Schlechta, K. and D. Makinson (1994). "Local and global metrics for the semantics of counterfactual conditionals." In: *Journal of Applied Non-Classical Logics* 4.2, pp. 129–140. DOI: 10.1080/11663081.1994.10510829 (cit. on p. 90).

Shanahan, M. (2016). "The Frame Problem." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University (cit. on p. 104).

Shannon, C. E. (1940). "A Symbolic Analysis of Relay and Switching Circuits." MA thesis. Massachusetts Institute of Technology. URL: https://dspace.mit.edu/handle/1721.1/11173 (cit. on p. 21).

Shapiro, L. and J. Beall (2021). "Curry's Paradox." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University (cit. on p. 6).

Sider, T. (2010). *Logic for Philosophy*. Oxford: Oxford University Press (cit. on pp. 4, 27, 29, 34, 41, 44 sq., 47).

Sørensen, M. H. and P. Urzyczyn (2006). *Lectures on the Curry-Howard Isomorphism*. Studies in Logic and the Foundation of Mathematics 149. Amsterdam: Elsevier (cit. on pp. 67, 69 sq.).

Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press (cit. on pp. 74, 89).

Stalnaker, R. C. (1976). "Propositions." In: *Issues in the Philosophy of Language: Proceedings of the 1972 Colloquium in Philosophy*. Ed. by A. F. MacKay and D. D. Merrill. New Haven and London: Yale University Press, pp. 79–91 (cit. on p. 25).

Starr, W. (2019). "Counterfactuals." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University (cit. on pp. 86, 90).

Stenning, K. and M. van Lambalgen (2008). *Human Reasoning and Cognitive Science*. A Bradford book. Cambridge, Massachusetts: MIT Press (cit. on p. 105).

Strasser, C. and G. A. Antonelli (2019). "Non-monotonic Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University (cit. on pp. 99, 106).

Tarski, A. (1956). "The Concept of Truth in Formalized Languages." In: *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Trans. by J. Woodger. First published in 1933 in Polish. Oxford: Clarendon Press, pp. 152–278 (cit. on p. 52).

Tichý, P. (1976). "A counterexample to the Stalnaker-Lewis analysis of counterfactuals." In: *Philosophical Studies* 29, pp. 271–273 (cit. on p. 94).

Troelstra, A. S. (1992). *Lectures on Linear Logic*. CSLI Lecture Notes 29. CSLI Publications (cit. on p. 67).

Van Atten, M. (2017). "The Development of Intuitionistic Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University (cit. on pp. 57, 68, 70).

Van Benthem, J. (2017). *Truth Maker Semantics and Modal Information Logic*. Tech. rep. Technical Notes (X) Series. X-2017-02. University of Amsterdam, ILLC. URL: https://eprints.illc.uva.nl/id/eprint/1590 (cit. on p. 80).

Van Rooij, R. (n.d.). "Self-referential truth." Course notes, University of Amsterdam (cit. on p. 125).

Veltman, F. (1985). "Logics for Conditionals." PhD thesis. Amsterdam: University of Amsterdam (cit. on pp. 94, 103).

— (2006). "Counterfactuals, the standard theory." Course notes, University of Amsterdam. URL: https://staff.fnwi.uva.nl/f.j.m.m.veltman/papers/Notes_Counterfactuals.pdf (cit. on pp. 91, 94 sqq., 106).

Williamson, T. (1999). "On the Structure of Higher-Order Vagueness." In: *Mind* 108.429, pp. 127–143 (cit. on p. 44).

— (2018). "Counterpossibles." In: *Topoi* 37, pp. 357–368 (cit. on p. 96).

Yanofsky, N. S. (2003). "A universal approach to self-referential paradoxes, incompleteness and fixed points." In: *Bulletin of Symbolic Logic* 9.3, pp. 362–386 (cit. on pp. 112, 115–118, 121, 125).

— (2022). *Theoretical Computer Science for the Working Category Theorist*. Cambridge Elements. Cambridge: Cambridge University Press (cit. on p. 115).

Index