Title
Paradox in Thought and Natural Language

Author
Jerzak, Ethan

Publication Date
2019

Peer reviewed|Thesis/dissertation

Paradox in Thought and Natural Language

By

Ethan J Jerzak

A dissertation submitted in partial satisfaction of the
requirements for the
degree of Doctor of Philosophy

in

Philosophy

in the

Graduate Division
of the
University of California, Berkeley

Committee in Charge

Professor John MacFarlane, Co-chair
Professor Seth Yalcin, Co-chair
Professor John Steel

Summer 2019

Paradox in Thought and Natural Language

This dissertation is entirely the original work of the author.

Abstract

Paradox in Thought and Natural Language

by

Ethan J Jerzak

Doctor of Philosophy in Philosophy
University of California, Berkeley

Professor John MacFarlane, co-chair
Professor Seth Yalcin, co-chair
Professor John Steel

Around 600BC, Epimenides, a Cretan apparently discontented with the honesty of his compatriots, lamented that all Cretans are liars. Together with a few innocent assumptions, well-entrenched principles of logic entail that Epimenides' lamentation cannot be true, and yet cannot be untrue—a flat contradiction. What's gone wrong? In this dissertation, I argue that the source of the problem has been misdiagnosed as one about language (especially formal languages). The problem runs deeper, and stems from the structure of thought itself.

The dissertation proceeds in two main stages. The first stage (Chapter 2) makes the case that that the intuitions that underlie the paradoxes come from natural languages, not from formal/mathematical ones. The Liar and related paradoxes are generally presented as constraints on the latter. Their lesson, the story goes, is that no formal theory strong enough to represent the primitive recursive functions can include a satisfactory truth predicate. I argue that it's our natural-language competence with the truth predicate that underlies our understanding of what 'satisfactory' means here, which shifts the focus of the project to natural language semantics. In this domain, it's tempting to think (and many have thought) that the problem with Epimenides' utterance is that it fails to express a proposition, and this failure explains

why we have trouble assigning it a truth-value. Or, perhaps it does express a proposition, but not the one that it seems to express. Or, perhaps it can express a proposition, but which proposition it expresses depends on context. I argue that all such responses fail, in part because they cannot make sense of related attitude attributions. I can believe or disbelieve Epimenides, which wouldn't be possible if his utterance didn't express the proposition it seems to express.

In the second stage, I argue that such paradoxes arise, not from the language/thought interface, but rather from thought itself. The first step in this argument concerns knowledge attributions (Chapter 3), where I develop and defend a novel solution to the Knower paradox. Then I move from attitude attributions to attitudes themselves (Chapter 4). Just as sentential truth and knowledge predicates gives rise to paradoxical sentences, seemingly innocent combinations of beliefs and desires give rise to paradoxical propositions—even when those beliefs and desires are not expressed in language. The possibility of such pathological combinations isn't accounted for by any extant theory of mental content, and, I argue, provides support for a non-classical theory. Finally (Chapter 5) I consider an objection to these putative combinations of desires. I introduce what I call *advisory* desire reports, which seem to exhibit the radically externalist behavior that the previous chapter rejects. I conclude by offering reasons to think that the availability of these readings does not undermine the case for non-classical accounts of attitudes.

Acknowledgments

There are many humans and institutions without which this dissertation could never have come to be.

To my parents and siblings gratitude is due for myriad reasons, but especially for their politely declining to inquire too scrupulously into what graduate school in philosophy entails and portends.

The greatest portion of intellectual debt I owe to the graduate community at Berkeley. Most of what I know about anything, I learned from arguments in 301 and at Wollheims. Particular (though by no means exhaustive) thanks to Adam Bradley, Melissa Fusco, Jim Hutchinson, Alex Kerr, Arc Koceruk, Richard Lawrence, Sven Neth, Emily Perry, Kirsten Pickering, Rachel Rudolph, Pia Schneider, and Umrao Sethi for particularly instructive yelling in this regard.

Many people gave invaluable comments on drafts and presentations of embryonic forms of what follows. Thanks to Chloé de Canson, Wes Holliday, Hannes Leitgeb, John MacFarlane, and Seth Yalcin for comments on drafts of Chapter 3. Thanks also to audiences at Berkeley, ESSLLI, and the MCMP for entertaining presentations about it. For Chapter 4, thanks to Wes Holliday, Arc Kocurek, John MacFarlane, Sven Neth, Rachel Rudoph, and Seth Yalcin for helpful discussions and comments, and to audiences in Berkeley, San Diego, London, Barcelona, and Oxford. And for Chapter 5, thanks to Chloé de Canson, Janice Dowell, Arc Kocurek, Hannes Leitgeb, John MacFarlane, Francois Reçanati, Rachel Rudolph, and Seth Yalcin for helpful draft commenting, to an anonymous reviewer at *The Journal of Philosophy* for helpful comments and challenges, and to audiences at UC–Berkeley, UCL, the University of Konstanz, the Institute Jean-Nicod, and the MCMP.

I'm very grateful to the DAAD, whose research grant allowed me to spend the 2015/16 academic year at the MCMP in Munich, and because of

whose two language grants I now speak passable German. Thanks to Hannes Leitgeb and the MCMP for agreeing to host me during this year; my time there was invaluable in fixing Chapter 3, and in writing the bulk of Chapter 5. And warm thanks to Catrin Campbell-Moore, Johannes Korbmacher, Hans-Christoph Kotzsch, Lavinia Picollo, Thomas Schindler, and Pia Schneider, who welcomed me into the MCMP community.

I’d also like to thank the Mellon Foundation and American Council of Learned Societies, from whom I received a dissertation completion fellowship for the 2017/18 academic year. During this time I published some of the more finished parts, and wrote Chapter 4.

Tons of advisers and mentors, both formal and informal, have helped me over the years to think and do philosophy better. Agnes Callard, Robert Pippin, and Jim Conant were all influential in kindling the desire to do philosophy. Without Tom Lockhart I might never have had the wherewithall to apply in the first place.

At Berkeley, Hannah Ginsborg and Barry Stroud were formidable first year seminar leaders, and collectively taught me the value of slowing down. Wes Holliday provided insight and guidance in the writing of Chapter 3. Seth Yalcin, despite sternly warning me in my first year not to write a dissertation on the present topic, was otherwise a model of encouragement. And John MacFarlane was extremely helpful, both through his meticulous comments on drafts, and in being most active in reminding me that I would eventually have to turn in a dissertation.

Two entities deserve special mention. First, the taco truck parked outside the Hotsy Totsy in Albany was a (sometimes alarmingly exclusive) source of nourishment during the final writing stages. And Schmidt’s Pub in Albany served as the location where the greatest number of words contained herein were actually penned. To these two establishments jointly is this dissertation dedicated.

Contents

Chapter 1

Introduction

The Liar paradox is one of the oldest problems in philosophy. Sometime around 600BC, Epimenides, a Cretian apparently discontented with the honesty of his compatriots, lamented that all Cretians are liars. Together with a few assumptions—that liars are those who only say false things, that as a matter of fact all other Cretians *are* liars, and that everything Epimenides had said until then was a lie—well-entrenched principles of logic entail that Epimenides' statement cannot be true, and yet cannot be untrue, a flat contradiction. And these assumptions are not necessary in contemporary versions of the paradox. Suppose Epimenides had simply said:

(1) The sentence I'm uttering right now isn't true. (Liar sentence)

Then (1) can't be true—if it were true, then it wouldn't be true, since it says that it isn't true. But it also can't be untrue, for if it were untrue, then (1) would describe the world accurately after all, and so be true. Contradictions are very good evidence that something has gone horribly wrong. The question for philosophers and logicians is, what?

A flurry of philosophical and mathematical research in recent decades has not settled on an answer. It has, however, characterized more precisely a space of possible answers. Contemporary reactions to the Liar paradox either:

- Place expressive limitations on languages, so that Epimenides' sentence isn't actually a legitimate sentence. (Tarski (1936))
- Reject the principle that, for any sentence φ, φ is true if and only if φ. (Kripke (1975))

- Reject some part of classical propositional logic. (Priest (1987), Field (2008))

This research has largely taken place in the context of formal languages, like those of arithmetic or set theory. This formality has been tremendously productive in characterizing the trade-offs involved in different treatments of the paradox. The upshot of all this abstract theorizing for the natural languages in which the Liar sentence originally arose, however, is far from clear. How should we think about these paradoxes from the perspective of everyday English, their original source? Should we look for empirical evidence that English mirrors one or another of these formal languages, or do natural languages present unique considerations that open space for more possibilities, or constraints, than pure formal logic? How should those formal theories be incorporated into the descriptively-minded semantics that linguists and philosophers of language have been developing for fragments of natural language? These are among the questions I address in this dissertation.

A related and largely unexplored question is about propositional attitudes like belief, knowledge, and desire. Just as paradoxical sentences can arise using the truth predicate, such sentences also arise in natural languages with knowledge, belief, and desire attributions. Thus the question: what would it be to *believe* the Liar sentence? Or to be in some paradoxical state of *desire* which, as a matter of logical necessity, both must and cannot be satisfied? The possibility of such pathological attributions isn't accounted for by any extant theory of propositional attitudes. These two problems, I show in this dissertation, are intimately related by the way we attribute and reason about propositional attitudes in natural language.

In this dissertation, I show how to adapt particular formal theories of truth to the project of natural language semantics. I also show how to develop plausible theories of intentional attitudes and their contents against this background. My main contentions are that, while propositions must be part of any satisfactory theory of truth in natural language, a brute appeal to them to solve the paradoxes is unpromising (Chapter 2). Instead, a non-classical account of logical consequence, which restricts certain classically valid laws of logic, is independently motivated by theorizing about the attribution of intentional attitudes like knowledge (Chapter 3). I bolster this argument by showing that these paradoxes can arise even at the purely propositional, non-linguistic level: thought alone is sufficient to generate self-referential contents, and therefore paradoxes (Chapter 4). The most promising case for this involves desire, and the strongest counterargument appeals to the information-sensitivity of desire attributions. I consider the information-sensitivity of desire attributions independently (Chapter 5), arguing for a novel relativist theory. I conclude by suggesting that this information-sensitivity does not suffice fully to disarm the desire paradox, and explore avenues for more detailed future work in this direction.

In what remains, I'll walk through an overview of the structure of my dissertation, pointing out work already done and contributions to the current literature.

1.1 Chapter 2: Liars, Propositions, and Contexts

Here is a rough but natural picture of how (at least some core parts of) natural language work. Language is a tool for communicating our thoughts. It's the job of declarative sentences to express the contents of these thoughts, which are often called propositions, and these propositions are true or false depending on what the world is like. It's in virtue of expressing these propositions that sentences can be true or false. For example, the sentences "snow is white" and "Schnee ist weiß" both express the same proposition, that snow is white, and whether those sentences are true depends on whether the corresponding proposition is true. If our world happens to be so arranged as to make snow white, then the proposition is true, and thus so are the sentences.

Given this picture of how declarative sentences attain their truth values, a natural first-pass diagnosis of the problem with the Liar sentence (1) is that it simply fails to express a proposition. There is no legitimate content/thought for the Liar sentence to express. This failure to express a proposition explains why the sentence exhibits paradoxical behavior, and absolves theorists of propositional attitudes from taking the paradoxes seriously. In this chapter, I examine the most influential no-proposition theory, due to Glanzberg (2001). He is a contextualist about truth, which means that (a) sentences have truth-values in virtue of expressing propositions with those truth-values, and (b) exactly which proposition a sentence expresses depends on the context in which it is asserted. He argues that there's a hidden context shift in the reasoning of the Liar paradox, which explains what goes wrong while saving classical logic.

I argue that Glanzberg's theory fails to achieve one of its major aims, which is to explain the unified inferential use we make of the truth predicate. I show that he does not avoid implausibly fragmenting the truth predicate when we reason with it, without independent motivation from linguistic data. Contextualist solutions must reject the inference from 'p' to 'p is true', accepting only the weaker rule from 'p' to 'p is true$_i$' for some contextually dependent i. In this respect, Glanzberg's propositional theory isn't any better than other theories which reject truth principles; they both considerably weaken the inferential power of the truth predicate. I conclude the chapter with two problems for this view. Adding propositions to the apparatus of truth-talk doesn't solve the paradox, at least not on its own.

1.2 Chapter 3: Non-Classical Knowledge

In this chapter, I examine more closely the relationship between paradoxes and propositions. The starting point is an independent role that propositions are taken to play: serving as the objects of intentional attitudes like belief, knowledge, and desire. I show how paradoxes structurally similar to the Liar paradox arise using these notions, taking knowledge as my primary case study. Drawing on the literature of the Knower paradox (Maitzen (1998), Cross (2001), Uzquiano (2004), Sainsbury (1995), Kaplan and Montague (1960)), I argue for modeling knowledge-talk in a non-classical language, analogous to a theory of truth developed by Hartry Field (2008). I compare this proposal to other theories of paradoxical attitudes, notably Caie (2012)'s argument that our beliefs about the Liar should inherit the same indeterminacy present in the Liar sentence itself. I show how my treatment of knowledge improves on his proposal, because it does not rely on controversial introspection principles, which say that you can never wrong about what you believe or desire. I also introduce a natural-language version of the Knower paradox, which, unlike the extant literature, makes the paradox not about the *ideal* knowability of sentences with mathematical content, but about the straightforward knowledge of actual agents reasoning with English sentences.

1.3 Chapter 4: Paradoxical Desires

While Chapter 3 treats knowledge *attributions*, there is an open question about the relationship between the state of knowing, and the way we talk about that state in natural language. It's open to argue that, while a non-classical theory perhaps best models the way we *talk* about mental states like knowledge, the fundamental nature of these mental states themselves doesn't require us to posit any paradoxical contents. Paradoxes do arise in thought, but only derivatively from the language we use to talk about and express those thoughts.

I think this view is mistaken. In Chapter 4, I present a paradoxical combination of desires. I show why it's paradoxical, and consider ways of responding to it. The paradox saddles us with an unappealing disjunction: either we reject the possibility of the case by placing surprising restrictions on what we can desire, or we revise some bit of classical logic. I argue that denying the possibility of the case is unmotivated on any reasonable way of thinking about mental content. So the best response is a non-classical one, according to which certain desires are neither determinately satisfied nor determinately not satisfied. Thus, theorizing about paradoxical propositional attitudes helps constrain the space of possibilities for adequate solutions to semantic paradoxes more generally.

1.4 Chapter 5: Two Ways to Want

My argument for the existence of paradoxical combinations of attitudes involves denying what I call radical externalism—the view, namely, that the content of someone's strongest desire can depend blankly on the desiderative states of someone else (even if the desirer is entirely unaware of them). However, certain uses of desire attributions seem to exhibit this kind of information sensitivity. This chapter, therefore, is solely dedicated to the question of how information factors into the content and truth value of desire reports. I present hitherto unexplored and unaccounted for uses of 'wants'. I call them **advisory** uses, on which information inaccessible to the desirer herself helps determine what it's true to say she wants. I show that extant theories by Stalnaker (1984), Heim (1992), and Levinson (2003) fail to predict it. I also show that they fail to predict true indicative conditionals with 'wants' in the consequent. I argue that these problems are related—intuitively valid reasoning with modus ponens on the basis of the conditionals in question results in unembedded advisory uses.

I consider two fixes, and end up endorsing a relativist semantics, according to which desire attributions express information-neutral propositions. The truth of a desire attribution, on the view I arrive at, depends on the state of information at the context of assessment. I compare 'wants' with 'ought', which exhibits similar unembedded and compositional behavior. Finally I sketch a pragmatic account of the purpose of desire attributions that explains why it made sense for them to evolve in this way.

1.5 CONCLUSION

Surprising consequences arise from the resulting picture. For one, states of desire and belief can be such that they stand in indeterminate relations, regarding satisfaction and truth, to the actual world. For another, there is no such thing as a completely general negation operation on these contents that returns *true* just in case the negated proposition fails to be true for *any* reason (including by lacking on a truth value). Finally, the link between suppositional reasoning and a commitment to conditionals is even looser than is commonly supposed. The argument from A to B can be logically valid while the conditional $A \rightarrow B$ fails to be true.

Chapter 2

Liars, Propositions, and Contexts

There are two compatible but distinct spirits in which to approach semantic paradoxes. On the one hand, we can take up the perspective of language builders. On this approach, pathological sentences like the Liar impose restrictions on how we can construct formal languages, much like physical laws impose restrictions on the kinds of stable buildings that we can build. We can use this material, but if we do, we can't support a building of a certain height; if we want to build to that height, we will need to use a different material. Similarly with the Liar: You can have, for example, a truth-predicate as part of your object language, but if you do, you have to amend classical logic, restrict seemingly innocuous inferences involving the truth predicate, or restrict your ability to name sentences at will. Or you can banish the truth predicate from your object language, in which case you can keep all of the classical logic and syntactic expressiveness that you want. On this very general approach—which is more or less the way the Liar is approached from the standpoint of formal logic—the challenge is to build the most robust formal language as possible, given the looming threat of semantic paradoxes.

There is another way to think about the Liar. We could take the perspective, not of language-builders, but rather as theorists of already existing natural languages. For at least *prima facie*, English contains all of the resources that make paradoxes inevitable. We shouldn't, I think, follow Tarski in concluding that English is just defective, and that ideal thinkers (on some interpretation of 'ideal') should abandon it in favor of one of the sanitized languages constructed in the above spirit. Indeed, it's not clear just how we'd go about doing that. All of the reasoning that led Tarski to this conclusion itself occurred in Polish. If Tarski can reason his way to dissatisfaction with Polish using inference rules from Polish itself, then, by his own lights, we need not throw away the baby in order to dispose of the bathwater. A semantic theory of English that entails that every sentence is a theorem should, for that very reason, be rejected. The corresponding metaphor for this approach would be: We come across a perfectly stable building, but one that, according to our current theories of physics and engineering, seems unable to stand. Yet clearly it does. Our task is to figure our *how* it stands, given what else we know.

In this chapter, I will be concerned with attempts by Burge, Glanzberg to give contextualist analyses of the paradoxes in natural languages. Burge and Glanzberg take different approaches: Burge thinks that the English truth-predicate has a hidden indexical parameter, and he develops a Tarski-inspired hierarchy to avoid inconsistency. Glanzberg rejects such a fragmentation, arguing that indexical behavior just isn't observed in ordinary uses of 'true' in English. He attempts to retain a unitary truth predicate by introducing propositions as the fundamental bearers of truth, and positing context-sensitivity in the domain of quantification over those propositions.

I argue that Glanzberg's proposal is best understood as providing metalinguistic truth-conditions for our truth-talk in English, rather than as a suggestion for modeling the inferences that we explicitly make about truth. I work out, as Glanzberg does not, exactly how his proposal is to be assimilated into the inferential system of English itself, and conclude that making this explicit removes the most substantial differences between his approach and Burge's. Glanzberg will have to fragment the truth predicate after all, at least, whenever we actually reason using it. In a way, Glanzberg acknowledges this, but he does not do so in a way that makes the relationship between his approach and those like Burge's perspicuous. Glanzberg's proposal is best understood as a way of *motivating* something like Burge's approach, rather than rejecting it. The only complaint about contextualist solutions from which Glanzberg, but not Burge, is immune, is that it is *ad hoc.* I'll conclude by raising two very general worries for any context-hierarchical interpretation of the English truth predicate: the arms-race problem, and propositional attitudes toward contingently paradoxical sentences.

2.1 A brief introduction to the Liar

I am not the first, nor will I be the last, to introduce the Liar from ground zero. For this, I beg patience; it will be helpful to fix notation, and to focus on the problem from the perspective of natural languages. Say, then, that you're taking a true/false test, and you come across the following question:

> T/F: This sentence is false.

You scratch your head, wondering whether to circle T or F. You reason:

> Suppose I circle T. For that sentence to be true is for it to get the world right; it is to tell things like they are. So in circling T, I am committing myself to whatever that sentence says. That sentence says that that sentence is false. What's false can't be true, so clearly I can't circle T. So I should circle F. But if I circle F, then I'm claiming that this sentence gets things wrong. Which is just to say, it's false. But that's just what the sentence says! So the sentence gets things right after all. Which is just to say, it's true.

At this point, you might refuse to circle anything, or you might circle both T and F, or you might tear up the test. All of these responses have analogues in the literature. Keep our test-taker's dilemma in mind; the above will serve as a canonical bit of locally intuitive but globally paradoxical English reasoning.

Here is the simplest formal language in which the test-taker's predicament can be modeled. Start out with a standard propositional language without quantification $\mathcal{L}_P$:

$$At := A_i$$
$$\varphi := At \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \rightarrow \varphi$$

Let's fix a standard classical semantics for this language. A model $\mathcal{M}$ consists of an interpretation function $v : At \rightarrow \{0, 1\}$. The semantic values for complex formulae are computed as follows:

$$[\![A_i]\!] = v(A_i)$$
$$[\![\neg\varphi]\!] = 1 - [\![\varphi]\!]$$
$$[\![\varphi \vee \psi]\!] = max([\![\varphi]\!], [\![\psi]\!])$$

$$\llbracket \varphi \wedge \psi \rrbracket = \textbf{\textit{min}}(\llbracket \varphi \rrbracket, \llbracket \psi \rrbracket)$$
$$\llbracket \varphi \rightarrow \psi \rrbracket = \textbf{\textit{max}}(1 - \llbracket \varphi \rrbracket, \llbracket \psi \rrbracket)$$

This language, with these semantic clauses, is safe from the Liar in an important sense: No restrictions on $\mathcal{M}$ are necessary. Any $\mathcal{M}$ will result in an interpretation that does not crash on complex formulae. This will not remain the case for the richer languages we'll consider.

To get the problem posed by the Liar into view, consider an enriched language $\mathcal{L}_{P+}$, which adds to $\mathcal{L}$ only singular terms s_i and a monadic predicate $T(x)$:

$$At := A_i$$
$$t := s_i$$
$$\varphi := At \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \rightarrow \varphi \mid T(t)$$

A model $\mathcal{M}$ for $\mathcal{L}_{P+}$ is a pair $\langle v, \mathit{ref} \rangle$ of an interpretation function $v : At \rightarrow \{0, 1\}$ and a naming function $\mathit{ref} : t \rightarrow \varphi$ for φ wffs of $\mathcal{L}_{P+}$. The controversy is what to say about the semantics and inferential behavior of T.

Maudlin (2004) distinguishes two ways to bring out the problem posed by the Liar paradox: the semantic way, and the inferential way. The **semantic version** of the paradox gets its bearings by asking directly: "What's the truth-value of *l*?" Our truth predicate seems to have the following meaning:

$$\llbracket T(t) \rrbracket = \llbracket \mathit{ref}(t) \rrbracket$$

This would validate an axiom schema in the object language:

$$T(\ulcorner \varphi \urcorner) \equiv \varphi \tag{T-Schema}$$

Here $\ulcorner \varphi \urcorner$ is a metalinguistic name for any term s_i of $\mathcal{L}_{P+}$ such that $\mathit{ref}(s_i) = \varphi$.

Recall that we did not impose any restrictions on *ref*. Therefore, here is a perfectly possible assignment of names to sentences:

$$\mathit{ref}(s_1) = \neg T(s_1) \tag{Liar}$$

s_1, that is, denotes the sentence $\neg T(s_1)$. Call any sentential name like this—that is, any name l such that $ref(l) = \neg T(l)$—a **Liar name** (relative to $\mathcal{M}$).[1]

On the semantic understanding of the Liar paradox, we've already generated the paradox. Applying the semantic clause for the truth predicate to a Liar name, we get:

$$[\![T(l)]\!] = [\![ref(t)]\!] = [\![\neg T(l)]\!] \tag{2.1}$$

We run immediately into a problem: $[\![T(l)]\!] = [\![\neg T(l)]\!] = 1 - [\![T(l)]\!]$. Therefore, we cannot assign l a semantic value, at least against the backdrop of the semantic framework we've formulated. If we assign it semantic value 1, we're forced to assign it 0; and if we assign it 0, we're forced to assign it 1. This is the semantic understanding of the paradox: we ask directly what semantic value we should assign l, and run into a problem given a naive semantics for the truth predicate.

The key point is that, on this approach, the problem is consistently to assign l a truth value. The actual reasoning about which truth values it should receive occurs in the metalanguage. We start by making observations about the meaning of the word 'true', and then use these observations to generate a semantic entry for 'T'. Metalinguistic reasoning about whether $T(l)$ should be assigned semantic value 1 or 0 is what generates the paradox.

There's a different way to present the problem posed by the Liar: what Maudlin calls the **inferential version**. Instead of focusing on the semantics of the English truth-predicate, we can look directly to the inferential structure of the predicate as it's used in English reasoning. For, when we want to find out whether to believe something, we can engage in inferences from other things that we believe, and we do so without necessarily engaging in sophisticated metalinguistic reflections about the meanings of words. Indeed, when we're giving a semantic theory for some fragment of English, we often take valid inferences as *data* that our theory should account for. What kind of data do we have for the English truth predicate? The following two rules have famously seemed constitutive of the meaning of the truth predicate:

[1] A slightly more fleshed out treatment of this way of generating self-reference is developed in the appendix to Chapter 3. The thing to note here is that self-reference is enforced directly by the model, not via any arithmetization of syntax. I'll also employ corner quotes to refer (metalinguistically) to the name of a formula—so, for example, $T(\ulcorner A_1 \urcorner)$ is shorthand for the least $T(s_i)$ such that $ref(s_i) = A_1$.

$$\begin{array}{r|ll} i & T(\ulcorner \varphi \urcorner) & \\ j & \varphi & \text{T-out } i \end{array}$$

and

$$\begin{array}{r|ll} i & \varphi & \\ j & T(\ulcorner \varphi \urcorner) & \text{T-in } i \end{array}$$

These are natural-deduction rules in the flavor of Fitch. T-out, for example, says that anytime a sentence of the form $T(\ulcorner \varphi \urcorner)$ appears on line i, then φ may be written on any line $j > i$ citing i with T-out as the justification.[2]

Something in the ballpark of these inference rules must be valid. We can see this simply by observing that we can't ever assert,

(1) Grass is green, but 'grass is green' isn't true.

or

(2) 'Grass is green' is true, but grass isn't green.

Sentences of that form, it seems, cannot be true, at least when it is clear that no shift in meaning or idiolect is involved. The incoherence of these assertions doesn't have anything in particular to do with grass or greenness, so we have strong evidence that both of these rules are valid inferences in English. *Whatever* semantics you cook up for the truth predicate, even ones that incorporate truth-value gaps, should validate these rules.

But, remember, $l := \neg T(l)$. Then we reason,

$$\begin{array}{|ll} \begin{array}{|l} T(l) \\ \hline \neg T(l) \end{array} & \begin{array}{l} \text{Hypothesis} \\ \text{T-out} \end{array} \\ \neg T(l) & \textit{Reductio} \\ T(l) & \text{T-in} \end{array}$$

This proof models exactly the steps made in the informal reasoning of our test-taker. We get a proof in the object language, from no premises, that l

[2] See Magnus et al. (2018) for exposition. Numbered lines and justification will be omitted in sufficiently short proofs with sufficiently obvious steps.

must be false, and true. We've *proved* a contradiction. Since there aren't any premises to haggle over, one of the rules must be wrong, or else *l* has somehow to be ill-formed, or else the contradiction has to be illusory, or else the contradiction must be true (and the logic non-classical).

On the semantic way of motivating the Liar, the problem was consistently to assign *l* some truth value. Our reasoning about which truth values it could or could not receive itself occurred in the metalanguage. On the inferential understanding, on the other hand, the problem is to explain why a proof, in the object language, that *l* is true (and therefore also false) is not valid. The semantic version is concerned with giving a semantic theory of the truth predicate; the inferential version is directly concerned, not with semantics, but with the deductive system. Of course, these two ways of motivating the paradox are intimately related. A semantics validate inference rules, and obviously valid inference rules call out for a semantic theory that validates them. The difference is only with where to start: with the semantics, or with the deductive system.

The inferential version of the Liar, Maudlin claims, is *prima facie* harder to excise than the semantic version when it comes to natural languages. For we can cook up sophisticated paradox-free semantics for the English truth predicate all day, but it will be of no use if our semantics doesn't validate obviously valid English rules (or at least gives some insight into why they're not valid). It's hard to motivate abandoning very intuitive rules of inference, ones that, if you look about in the world at large, are commonly used and accepted. Therefore, it is incumbent upon any solution to the Liar paradox that it motivate abandoning one or another of these rules, in ways that don't make doing so seem like an *ad hoc* solution, cooked up solely for the purpose of blocking semantic paradoxes. The holy grail would be to find some linguistic evidence, from outside the fairly obscure context of the paradoxes, that motivates a solution restricting these inference rules.

2.2 Intuitive responses

I will now leave behind these preliminary observations, although they will become important later on. In this section, I'll sketch the most intuitive responses to the Liar, and examine their shortcomings. For, since my question is not, "What's the most powerful language that does not fall prey to semantic paradoxes?", but rather, "How does *English* work, such that

Liar-type reasoning does not commit us to inconsistency?", considerations of intuitiveness take on extra weight. Out of the failure of these intuitive responses will arise context-based approaches.

2.3 Banning self-reference

By far the most common response to the Liar one hears from non-experts is: "What kind of sentence is 'This sentence is false'? Proper sentences can't refer to themselves, they can only refer to things and perhaps to other sentences." Many have wrestled with the intuition that the Liar sentence is somehow defective, and that the paradox must rely on some nefarious philosopher's trick. Perhaps the trick was that the philosopher constructed what only *seems* to be a well-formed sentence, but which, due to illicit self-reference, actually fails to be syntactically well-formed.

One problem with this response is that it just isn't true of English. For example, "This sentence has five words" is straightforwardly true. Perhaps English just bans self-reference when semantic, rather than syntactic, predicates are involved? Sadly, banning explicitly self-referential sentences when semantic predicates are involved won't help with blocking the paradox, because of contingently paradoxical sentences. Indeed, contingently paradoxical sentences are the real test-cases for a theory of the English semantic predicate, because these sentences show that, not only can paradoxes arise in artificially constructed contexts (like test-taking), but that they can arise, albeit rarely, in the course of making responsible statements about the truth-values of other utterances. And, we'll see in Chapter 3, contingent types of paradoxes in other domains will serve an important argumentative role.

2.3.1 Contingently paradoxical sentences (Kripke (1975))

Surely you can understand the sentence: "Don't listen to your father; everything he says is a lie." And surely you also understand the sentence: "Listen to your mother; she's always right." If you understand these sentences, you're in the mire. For (simplifying a bit by removing the quantifications) these together have the form:

$s_1 : \neg T(s_2)$
$s_2 : T(s_1)$

We'll have the same semantic and inferential problems for this set of sentences as arose for the simple Liar. If s_1 is true, then s_2 is false; but s_2 just said that s_1 was true, so s_2 would have to be true. Contradiction. On the other hand, if s_1 is false, then s_2 is true; but s_2 just said that s_1 was true, so it would follow that s_1 is true after all , contradicting our supposition that s_1 is false.

Again, we can weasel our way out of semantic versions by assigning both of these sentences some junk truth value (call it $\frac{1}{2}$) rather than 0 or 1; but there are corresponding deductive versions, modeling line-for-line this informal reasoning, that seem to force us to give up some natural rules of inference. In any case, this shows that, whatever is the real problem with self-referential (sets of) sentences containing truth predicates, it's not that they aren't proper sentences. Most of the time, both s_1 and s_2 are unproblematic; it's an empirical question whether s_1 and s_2 together give rise to paradox. Whatever is wrong with them, they're not syntactically ill-formed.

2.4 Failure to express a proposition

These contingently paradoxical sentences show that English has the syntactical resources to construct sentences involving the semantic predicate that can lead to paradox if circumstances are sufficiently unfavorable. But we can try to recover a kernel of wisdom from the no-sentence approach, by taking the following tack. First, what exactly are we *believing* if we accept *l*? It seems clear that the extra-linguistic world stays exactly as it is whether or not we assent to *l*. There's no thought about the world that is at stake in *l*. So perhaps *l*, though a genuine, syntactically well-formed sentence, fails to express a truth-evaluable claim at all. For further motivation, compare the Liar sentence to other infelicitous utterances:

(3) I promise that I won't keep this promise.

(4) I command you not to obey this command.

We could try to motivate a similar paradox: Say that you keep the promise; then it follows that you didn't keep the promise, because you promised not to keep it. If you don't keep it, then you keep it after all. Do these reflections suggest that our concepts of promising or commanding are incoherent? No: They tells us that there is no such promise as (3), and no such command as (4).

Things might look better with contingent versions of these paradoxes. For example:

(5) A: I command you to obey all of B's commands.
B: I command you to disobey all of A's commands.

As with the contingent Liar, it will be hard to say what it would be to obey A's command, once B utters their (prima facie) command. The no-command response would have it that whether B's utterance counts as a command depends on what A has commanded.

This may strike one as odd. To know what *kind* of speech act B is making, we have to know a lot about what the world entirely external to B is like. Admittedly, this in itself may not be so weird. If one happens upon a theater play and hears someone utter the words, "Help, Marjorie is dead!", one may not be aware whether this is a proper assertion and request, or just one of the mock assertions and requests that one hears in the course of a theater play.

But in cases like this, the difference in type of speech act, though dependent on the surrounding context, clearly trace back in some way to the intentions and beliefs of the conversational participants. In (5), B intends to be issuing a proper command, and nothing about this intention changes depending on what A commands.

Still, there is some pull towards the idea that B's "command" is somehow defective, and therefore not a full-fledged command, because of what A has commanded. We can see this by asking what B would say if they were informed about what A has previously commanded. It would be odd for B to stick by what they said earlier, or to complain that their command was not followed, or to rejoice that it was followed, no matter what actions were performed. This suggests that no full-fledged command was issued, despite B's intentions.

What would a similar no-proposition view look like in the case of the Liar? On such a view, the Liar really amounts to the 'claim':

(6) The thought that this sentence expresses is false.

We can try the same sort of response: There is no such thought. Nothing syntactic is wrong with (6); what's wrong is that it fails to *say* anything. We cannot believe it or disbelieve it, because it is not a believable.

One benefit of this approach is that it seems to unify two domains in which people have argued for truth-value gaps. Many have thought that semantic paradoxes motivate truth-value gaps. But that hasn't been the only motivation: Frege held that cases of presupposition failure, like sentences involving non-denoting names, actually have no truth value, because they express no thought: they contain a term that does not denote (though it may have a sense). Such sentences express only mock thoughts, just as stage thunder merely appears to be thunder. On the no-proposition approach, the Liar is equivalent to a sentence which presupposes the existence of a thought for that sentence to express; like the case of empty names, this presupposition is false, and therefore the sentence lacks a truth-value. The no-proposition approach to the Liar would explain how these two motivations for truth-value gaps are actually quite related.

A few notes on terminology. I'll use 'thought' and 'proposition' more or less interchangeably. I want to remain as agnostic as possible on what these are, metaphysically speaking. I mean something like: 'The *content* of an utterance' or 'the truth-evaluable claim that an utterance makes'. This is a fairly intuitive idea. If on Monday you say, "It's raining today" and on Tuesday you say "It rained yesterday," these are just two different ways of saying the exact same thing. The same empirical fact makes both of those sentences true. These sentences therefore express the same proposition.[3] The problem with the Liar, on the view I'm trying to motivate, is that it fails to express *any* such proposition.

2.4.1 Problems for a straight no-proposition view

The above line is attractive as (6) is formulated. But we can push this response in directions that undermine its initial plausibility. First, let's make a new Liar sentence:

(7) The sentence labeled (7) does not express a true proposition.

Let's try to run the no-proposition line: "The problem with (7) is that it doesn't express a proposition." But wait: If it doesn't express *any* proposition, doesn't it follow that, in particular, it doesn't express a true proposition? And

[3] This is too coarse a way of individuating propositions in general—it would, for example, entail that there was just one mathematical proposition. But this commitment isn't necessary for what follows. The no-propositions views explored below don't assume much about the structure of propositions, and instead just ask about the prospects of a theory that deny that the Liar expresses one.

isn't that *exactly* what (7) says? Someone advocating the no-proposition view will have to maintain: "No. That's what it *would* say, if it actually expressed a proposition. But, on pain of paradox, it can't. So really (7) doesn't say anything, even if it *appears* to say that it doesn't express a true proposition."

But this looks hopeless. Consider the following two statements side-by-side:

> (7): The sentence labeled (7) does not express a true proposition.
> Theorist: The sentence labeled (7) does not express a true proposition.

Each of the words in these sentences has the same meaning (there's no bank-bank type equivocation), and they're put together in the same ways. Yet, an advocate of the no-proposition approach will have to say, (7) does not express a proposition, but the sentence uttered by the theorist does—indeed, it expresses a true proposition.

This suggests a difference between the self-refuting promise and the Liar sentence. For, if we try similarly to rephrase (3): "This sentence does not express a promise with the content that I won't keep that promise", there's no paradox: that sentence is simply true, because that sentence doesn't amount to a promise. If we rephrase it as "This sentence expresses a promise with the content that I won't keep that promise", it's just false: No it doesn't, because there is no such promise. With the Liar, things are harder.

We have two sentences with the same words in the same order, and no obvious equivocation in meaning between words, and yet the sentences have two different semantic values (none at all, or what we're representing as $\frac{1}{2}$, and 1, respectively). This observation has led many to suspect that there is some sort of context shift at play. It's actually not uncommon in natural language to find two tokens of the same sentence with different truth values. The following two sentences,

(8) This car is loud.
This car isn't loud.

can both be true if the referent of 'this car' or the meaning of 'loud' has changed based on context. Similarly, perhaps (7) and the theorist's diagnosis of what's wrong with (7) can unproblematically have different truth values. Perhaps there has been a hidden context shift between them.

2.5 Two kinds of context-based approaches

There's a lot of intuitive plausibility in a no-proposition analysis of the English semantic paradoxes. Of course, making such a view work will require telling a story about what propositions are and how they relate to sentences. But even if such a story can be told, the above line shows that any such view must account for the difference between (7) itself and the theorist's diagnosis of (7). Context is a natural place to look, for historically, context has been introduced in semantics to explain how different utterances of the same sentence can have different truth values on different occasions.

Those who have taken this line, however, have disagreed on the exact nature of the context-sensitivity involved in the Liar. Burge (1979b) thinks that the English truth predicate itself has a hidden indexical element, and that that uses of it on particular occasions are indexed (at the level of syntax) to particular levels on Tarski's hierarchy. He thinks that once we analyze the structure of our truth predicates in this way, we'll see that we don't need to posit propositions to do any explanatory work after all.

Glanzberg (2004) rejects Burge's approach, claiming that it implausibly posits a wild proliferation of different truth predicates at the syntactic level. Instead, he attempts to retain a single truth predicate by introducing propositions as the fundamental truth-bearers. When we say that a sentence is true, what we mean is that the proposition it expresses is true. He then claims that the problem with the Liar inference is that it hides context-sensitivity in the domain of quantification over propositions. (His hierarchical analysis of context is more Kripke-like than Burge's Tarski-inspired approach, but that difference won't much matter here.) I'll briefly introduce the idea behind Burge's account, and then cite Glanzberg's reasons for criticizing it. Then I'll argue that Glanzberg's proposal, properly understood, does not escape from fragmenting the English truth predicate for many practical purposes (the ones where we reason about truth using T-in).

2.6 Indexical truth predicate (Burge)

Burge (1979b) gives what is probably the most well-known contextualist analysis of semantic paradoxes. He reasons: Say that we take our initial proof of l and $\neg l$ to show that l must lack standard truth-conditions. (Whether we want to analyze this as a failure to express a proposition is, for the moment, another matter.) Since l lacks standard truth conditions, it follows in particular that l isn't true; that is, the sentence $\neg T(l)$ *is* true. But $l :=$ $\neg T(l)$. So l must be true after all. What we want, to prevent a contradiction, is to explain the difference between the reasoner's derivation of l and l itself, which look identical. Burge:

> [This reasoning] seems to involve no change in the grammar or linguistic meaning of the expressions involved. This suggests that the shifts in evaluation should be explained in pragmatic terms. Since there is a shift from saying that the relevant sentence is not true to saying that the same sentence *is* true—a shift in truth value without a change of meaning—there is an indexical element at work. The indexicality is most plausibly attributed to the truth predicate. (Burge (1979b), p. 179)

In rough outline, Burge proposes interpreting the truth predicate in l itself as T_i for some i, and interpreting the reasoner's derivation of l as using T_j. There is no explicit contradiction in asserting $\neg T_i(l)$ and $T_j(l)$. Burge then gives three different proposals for modeling the structure of our truth predicate. All are Tarski-inspired hierarchies, where you start with atomic, non-truth-involving sentences, and recursively construct increasingly powerful truth predicates, each of which can include only the lower-level ones in its extension. The main difference that Tarski's hierarchy posits infinitely many increasingly rich metalanguages, each of which contains the semantic predicate of the language below it. Burge's proposal posits a single indexically sensitive truth predicate in one language. So when we interpret the word 'true' in English, we must assign it some value on this hierarchy. So, Burge thinks, you can have an apparatus of propositions if you want, but it needn't be invoked to do any explanatory work in blocking the paradox.

2.7 Indexical domain of quantification over propositions (Glanzberg)

Glanzberg agrees with Burge, for basically the same reasons, that there must be a context-shift between l and the theorist's diagnosis of what's wrong with l. But he rejects an approach that fragments the truth predicate. His argument is worth quoting in its entirety—for, I will argue, his proposal does not avoid fragmenting the truth predicate that we actually reason with in English. Glanzberg:

> A more common idea [than mine] is to suppose that the truth predicate itself contains a hidden indexical component. Let me briefly note that I do not think this is a promising option. It is a commonly voiced objection to it that we simply to not intuitively see such an indexical element in our ordinary truth predicate, expressed by the ordinary term 'true'. I believe this line of argument can be bolstered. If there were such a hidden indexical, it would behave as other implicit parameters do. In the case of a gradable adjective, for instance, we can see the hidden comparison class at work when we *bind* the hidden variable, as in:
>
> Most species S have members that are small for S.
>
> We see no such behavior with the truth predicate. (Glanzberg (2004), p. 30)

Glanzberg's idea is this. When we attribute truth or falsity to our sentences, we're really attributing truth or falsity to the thoughts (or propositions) expressed by those sentences. So when we say " 's' is true", what we really mean is: "There is a true thought that 's' expresses." That is:

$$T(s) := (\exists p)[Exp(s,p) \& T(p)] \tag{2.2}$$

Glanzberg then gives us some basic principles for reasoning about truth. Those are:

$$Exp(\ulcorner \varphi \urcorner, p) \rightarrow (T(p) \leftrightarrow \varphi) \tag{T-Exp}$$

$$(Exp(s,p) \& Exp(s,q)) \rightarrow p = q \tag{U-Exp}$$

$$p = q \rightarrow [T(p) \leftrightarrow T(q)] \tag{T-Id}$$

Then, Glanzberg generates the paradox thus. Our Liar sentence should actually read:

$l : \neg(\exists p)[Exp(l,p) \& T(p)]$

And we reason using the above rules. Informally, it goes like this: Suppose l expresses a proposition p. If p is true, then p is false. If p is false, then p is true. Hence, l does not express a proposition. But then, it follows that, in particular, l doesn't express a true proposition. But that's just l. So l is true. But to be true is to express a true proposition. Paradox. (His formal proof requires all of his truth-rules, plus if-introduction, reductio, conjunction introduction, and the standard rules involving the quantifiers.)[4]

How to block the paradox, according to Glanzberg? His proof has two conclusions:

(A): $\neg(\exists p) Exp(l,p)$
(B): l

Glanzberg refuses to consider giving up any of the rules of inference that led to these results: "As both (A) and (B) are the results of sound proofs, they must both be true." But given that he has also rejected fragmenting the truth predicate (and, presumably, the Exp relation), where is the context sensitivity? Glanzberg:

> I maintain we have to see a context shift between (A) and (B) affecting l. We have a proof, based on solid principles, so its conclusions had better be correct. If there is no context shift, then we have a genuine contradiction, so there must be a context shift…The only locus for context dependence [given that there are no standard indexicals], is the domain of the quantifier $\exists p$. (Glanzberg (2004), p. 34)

Glanzberg goes on to give a fairly sophisticated formal model of context, which shows how (A) and (B) aren't inconsistent after all. I won't get into the details, which are quite technical, but the basic idea is to interpret the existential quantifier in (B) as ranging over strictly more propositions than that of (A). So what we really have is something like:

(A): $\neg(\exists_0 p) Exp(l,p)$
(B): $l = (\exists_1 p)[Exp(l,p) \& T(p)]$

[4] Glanzberg's formal proof can be found on page 33 of his 2004.

These are both true. The proposition that satisfies (B) simply isn't in the domain of $\exists_0$, although it is in that of $\exists_1$. When we asserted (A), we had to 'step back', and observe something like 'that's just what *l* says'. This 'stepping back', says Glanzberg, expands the domain of truth-conditions available for us to quantify over. And, of course, with continued iterations of the same argument, we'll get a sequence of seeming contradictions, which will, on Glanzberg's account, receive increasingly rich existential quantifiers. The Liar sentence will come out true on the odd indices, and false on the even indices. Glanzberg spends much of his paper developing a sophisticated formal model of context that explains how asserting things like (A) can continually expand the domain of truth conditions for sentences like (B). But the exact structure of his formal model isn't crucial for much of what follows.

2.8 Whither T-out and T-in?

It's not immediately clear how to understand Glanzberg's diagnosis. He claims to be specifically concerned with the paradox as it arises in natural languages, but when he gets around to stating his version of it, the proof involves propositions, expression relations, and quantification—none of which elements obviously appears in our ordinary English reasoning about truth. We *can* say, "The sentence that you uttered didn't express a true proposition in the context in which you uttered it", but I doubt that anyone unschooled in a fair amount of philosophy of language would put it that way. We say simply, "That's false!" or "Everything you've said about my supposed affair with the countess is false." The first has no quantification; and insofar as there's a quantifier in the second, it isn't completely clear whether it ranges over sentences or propositions.[5] And no lexical item corresponding to 'Exp' appears in either.

[5] One way to get an intuitive handle on this question is to ask about intuitions involving counting the objects of truth. If someone utters "John greeted Mary" and "Mary was greeted by John", and John indeed greeted Mary, how many true things did they say? Two answers seem acceptable: one (namely the proposition that John greeted Mary), formulated in two different ways, and two (the number of true sentences that they uttered). However, as the conversation grows more complex and the distinct propositions get harder to count, it becomes more natural to default to counting sentences. For example, in a conversation involving morning stars and evening stars, where it's less intuitively clear how to individuate propositions, sentences seem like a natural fall-back.

Now, it might be that the best way to analyze what we *mean* by such utterances will involve quantifying over propositions and relating them to sentences via something like 'Exp'. But if this is what's going on, it's not going on *explicitly* at the level of English syntax. It's going on 'under the hood', so to speak, in a metalanguage. Therefore, I submit that Glanzberg's proposal is most plausible when it's understood as providing truth conditions for our truth-talk in a metalanguage of English. He's giving a truth-conditional semantics for sentences involving the English truth-predicate.

I should note that I'm not entirely sure whether this aligns with Glanzberg's self-understanding. For one thing, the word 'metalanguage' does not appear in his paper, save in a few endnotes. For another, T-Exp allows us to go back and forth between sentences involving T, Exp, and propositions, and regular English sentences without those elements. On the interpretation I'm advocating, this can seem like a category error: Glanzberg's Liar inference would go back and forth between sentences in a metalanguage and sentences in the object language of which it's a metalanguage. But I submit that we must understand Glanzberg in this way, if his proposal is to have any plausibility. His truth-rules are fairly baroque, and I can't imagine that anything like T-Exp has ever appeared explicitly in English reasoning. If it's a part of our reasoning at all, it must be doing its work under the hood. It must be part of our semantic theory of the English truth predicate, not part of a serious proposal for modeling the valid English-language arguments that actually appear in the wild.

Thus, to revisit a distinction made above, Glanzberg is treating the paradox in the semantical way, not the inferential way. Though he does provide a 'proof', which suggests that he is concerned with rules of inference, this proof does not model intuitive English reasoning (like that of our test-taker). Furthermore, Glanzberg considers it beyond the pale to abandon any of these jointly troublesome inference rules. He just wants to show that they don't commit us to inconsistency. His concern is with the semantics of English, not directly with its deductive system.[6] For these reasons, then, we should understand his 'proof' as modeling reasoning that occurs in a metalanguage. But then, it's natural to ask: How exactly is Glanzberg's

[6] It is debatable whether English even *has* a deductive system, in the sense of syntactically specifiable rules that correspond to some well-defined notion of a formally valid argument. For these purposes I assume that there are at least some arguments that are formally valid in English, but this is not universally accepted. See, for instance, Quine (1951) and Russell (2018) for skepticism on this front.

proposal supposed to account for the inferences that we explicitly *do* make regarding truth (exemplified by our test-taker)? If Glanzberg is right about the semantics of our truth-talk, what does that imply for the deductive system of the language that's ultimately at issue?

As a first step toward answering this question, let's make Glanzberg's proposal more explicitly geared toward this understanding of it. He argues that we should give something like the following semantics for the English truth-predicate as applied to sentences:

$$[\![T(s)]\!] = 1 \text{ iff } (\exists p)[Exp(s,p) \& T(p)]$$

And recall that Glanzberg accepts all of his rules about truth. Clearly these play a role analogous to that of T-out and T-in. Do Glanzberg's rules governing truth in the metalanguage validate T-out and T-in in its object language? Recall,

$$Exp(\ulcorner \varphi \urcorner, p) \rightarrow (T(p) \leftrightarrow \varphi) \qquad \text{(T-Exp)}$$

$$(Exp(s,p) \& Exp(s,q)) \rightarrow p = q \qquad \text{(U-Exp)}$$

$$p = q \rightarrow [T(p) \leftrightarrow T(q)] \qquad \text{(T-Id)}$$

First, we can observe that T-out is straightforwardly valid: Suppose that $T(\ulcorner \varphi \urcorner)$. That means that there's a true proposition that $\ulcorner \varphi \urcorner$ expresses; and by T-Exp, φ follows. So T-out is valid.

What about T-in? One tempting way to interpret T-Exp is as a way of *blocking* T-in. We only get to invoke T-in if we already know that s expresses a proposition. But Glanzberg seems to think that we can assert s truthfully only if it expresses a proposition. ("The truth of l seems to require there to be a proposition for l to express." [Glanzberg (2004), p. 34]) If that's so, then an unrestricted version of T-in would be valid after all.

So, when it comes to the actual language that we all explicitly use and love, Glanzberg needs to follow basically the same approach as Burge. He accepts *all* of the inference rules that the test-taker used in his reasoning. He must therefore think that we have true instances of $T(l)$ and $\neg T(l)$. The only way for this to happen, by his lights, is for the truth predicate to shift extension based on context. Its shift in extension, according to Glanzberg, is derivative from an expansion of the domain of quantification over propositions in the metalanguage; but nonetheless, to avoid contradiction, its extension must so shift. And so, the truth predicate in English, applied to sentences, must be indexically sensitive after all.

To make it clear just how important it is that we keep an index on the truth predicate in our actual reasoning about truth, consider Curry’s paradox. Curry’s paradox is without a doubt the best semantic paradox. It shows directly how an explosion principle follows from only the T-inferences, *modus ponens*, and if-introduction. I’ll give it first in English, and then in Fitch-ese.

> Consider the sentence *c* that reads, “If ‘*c*’ is true, then your mother was a hamster”. Now, suppose that ‘*c*’ is true. ‘*c*’ just says that, if it’s true, then your mother was a hamster, so it would follow that your mother was a hamster. But we just derived “Your mother was a hamster” from supposing " ‘*c*’ is true“, so surely, *if* ‘*c*’ is true, *then* your mother was a hamster. But that’s just what ‘*c*’ says! So ‘*c*’ is true. So your mother was a hamster.

In Fitch-ese: Consider the sentence, $c := T(c) \supset \bot$. We can reason:

1	$\quad T(c)$	Hypothesis
2	$\quad T(c) \supset \bot$	T-out
3	$\quad \bot$	*Modus Ponens*
4	$T(c) \supset \bot$	$\supset$ Introduction
5	$T(c)$	T-in
6	$\bot$	*Modus Ponens*

How do we deal with Curry’s paradox, on Glanzberg’s analysis? Remember, Glanzberg rejects modifying or restricting classical logic out of hand, and so he has to find a way to domesticate this reasoning without concluding that there’s some context in which $\bot$ holds—for surely there’s *no* context in which your mother was a hamster.

But it’s fairly clear how Glanzberg’s analysis will go: (5) and (6) are both true in the context of assertion, but nonetheless we cannot apply *modus ponens*. For, when we invoked T-in, we ‘stepped back’ and observed, ‘That’s just what’*c*’ says’. Thus, when we go to the Glanzbergian metalanguage, we’ll find something like:

5: $\exists_0 p[Exp(c,p) \& T(p)] \supset \bot$
6: $\exists_1 p[Exp(c,p) \& T(p)]$

There is indeed a true proposition in the domain of $\exists_1$ that *c* expresses; but that proposition does not fall within the domain of $\exists_0$. Thus the antecedent of (5) is false, even though (6) is true.

The crucial point is that we must keep track of this when we reason about the truth of our sentences. The only way to do this is to posit an indexical parameter on the truth predicate in English whenever it's applied to sentences. T-out will still be straightforwardly valid, but when we invoke T-in, we'll need to make explicit what index we're associating with T. For, T-in is the inference rule that models the 'stepping back' move that Glanzberg thinks is liable surreptitiously to shift the context. So if we don't keep track of this in our statement of T-in, we'll have rules that lead, not only to outright contradictions, but to explosion (in the sense that every sentence of the language will end up being provable). Our actual statement of T-in has to be:

$$\begin{array}{|l} \varphi \\ T_i(\ulcorner \varphi \urcorner) \qquad \text{For some contextually-dependent } i. \end{array}$$

The indexical parameter on the truth predicate will correspond to that on the existential quantifier in the metalanguage. The English truth predicate, applied, as it usually is, to sentences, can't get *exactly* the analysis it got above, for failing to specify its place on the hierarchy of context led directly to Curry's paradox. Applied to sentences, the truth predicate means something like:

$$[\![T_i(\ulcorner \varphi \urcorner)]\!] = 1 \text{ iff } (\exists_i p)[Exp(\ulcorner \varphi \urcorner, p) \& T(p)]$$

Therefore, when it comes to modeling the deductive system of English, Glanzberg's approach more or less matches Burge's in spirit, if not in detail. Glanzberg might provide a motivation for fragmenting the truth predicate when we reason with it, and giving a more systematic underlying mechanism, but he does not escape doing so.

Now, it's not clear that this counts as a substantial objection to Glanzberg's project. He states at the beginning of his paper that his main motivation is to avoid the suggestion that a broadly contextual-hierarchical solution to the natural-language Liar is *ad hoc*. And he admits that what he calls 'internal' semantic relations—*roughly* what I've called the truth-predicate applied to English sentences—can shift in extension as context shifts.

However, I do wish to urge that the way he seems to conceive of his project in relation to other context-based approaches is misleading. In assessing the merits of his proposal, he says, "We do not start by positing a hierarchy of truth relations, as some hierarchical approaches do. Nor do we posit an index on the truth predicate (an approach I rejected)" (p. 78). As we've seen, this isn't quite right. It's true that he doesn't *start* by positing a hierarchy of truth relations, but he does end up there after reflecting on expanded domains of quantification. And his approach is only plausible if there *is* an index on the truth predicate when applied to sentences, in at least those cases in which T-in is invoked. Here are his remarks about the kind of indexicality he finds in the English word 'true':

> As it is driven by context dependence, the hierarchy I propose does not posit a lexical ambiguity of the truth relation. Though ultimately an internal truth relation is used in interpretation, this no more indicates a lexical ambiguity of the truth relation than [the denotation of 'that' changing based on context]. The meaning of the truth relation remains constant, just as the meaning of 'that' remains constant. But in both cases, when looking across context shifts, we have to be careful to reconstruct the right context-dependent value. (Glanzberg (2004), p. 79)

So, for Glanzberg, the merit of his approach lies in the fact that there's not an explicit lexical ambiguity; it's not as though 'true' must be assigned any number of unrelated predicates based on context. But this is basically what Burge thinks *he's* doing! Here's Burge defending himself against charges of 'fragmenting' one concept into many:

> What of the univicality criticism of Tarski?... In natural language there is a *single* indexical predicate. We represent this predicate by the schematic predicate expression $true_i$. This expression may in particular contexts be filled out by any of an unlimited number of numerical subscripts. Any one of the resulting predicates (formally, there are infinitely many) may represent a particular occurrence of 'true' in a context in which its application is fixed. Thus the numerals substituted for '*i*' mark *not* new predicate constants, but contextual applications of the indexical 'true'....[In

> this sense,] 'true' has a single meaning. (Burge (1979b), p. 191, my emphasis)

Both Burge and Glanzberg agree that there's not an explicit lexical ambiguity in the English 'true'. But it *is* indexically sensitive, in the same way as 'that' is indexically sensitive. When it comes to interpreting the extension of the predicate 'true', applied, as it usually is, to sentences, Burge and Glanzberg do not differ in their general approach (although the exact structures of the hierarchies they construct do differ somewhat).

2.9 Two problems for this approach

So much for this interpretative point. What we really want to know is: How plausible is this general approach? I'll conclude by raising two worries for the contextualist analysis. These worries are not about the motivation for contextualism about truth; they are about the limitations contextualism imposes on the thoughts we are able to express in English. Perhaps we should accept these limitations as unavoidable. But a response that succeeded equally in blocking the paradox and did not come with these limitations would be preferable.

2.9.1 Arms-races

The arms-race problem is one that Burge addresses head-on, but Glanzberg does not mention. The idea comes from Kripke's contingently paradoxical sentences. Suppose that you think that Jones is extremely unreliable, and never says anything true. And Jones thinks that you are wise, and would never say anything false. Leave aside for a moment the epistemic merits of either of these claims. The question is: Can both you and Jones express the thoughts that you want to express?

Let's provide a contextualist analysis of both of your statements. Both Burge and Glanzberg must interpret the 'true' and 'false' above as limited in extension to a place on a hierarchy of some sort. Whenever we see the word 'true' applied to English sentences (as it is in these examples), even Glanzberg has to tell us what level on the context-hierarchy it corresponds to. But which level should we pick? We have something that looks like:

You: $s_1 : (\forall s)(\mathit{JonesAsserts}(s) \rightarrow \neg T_i(s))$
Jones: $s_2 : (\forall s)(\mathit{YouAssert}(s) \rightarrow T_j(s))$

On Burge's proposal, either someone gets the short shrift, or else neither of you can be interpreted as saying exactly what you clearly mean. For on any of Burge's constructions, we can never see anything of the form $T_i(\ulcorner \ldots T_j \ldots \urcorner)$ if $j \geq i$. Sentences such as these are syntactically ill-formed; each T_i is specifically constructed to range over sentences with truth predicates below i on the hierarchy. So if $i = j$ above, then s_1 says ***absolutely nothing*** about s_2, and vice-versa. If $i > j$, then s_1 *does* assert the falsity of s_2, but s_2 says absolutely nothing about s_1. Burge thinks that the only plausible thing to do here is to bite the bullet, and make $i = j$ unless context obviously dictates otherwise. And so s_1 says nothing about s_2, and s_2 nothing about s_1. Perhaps it's worth biting this bullet. But a bullet it remains.

Another thing to point out is that, when I don't remember exactly how many semantic predicates so-and-so has embedded within his utterances, but I still want to express my disagreement, I'm under intense pressure to pick as high an i as possible when I assert that everything he says is not true$_i$. Otherwise I leave open the possibility that I won't have said something about all of the utterances that I wanted to say something about. And as time goes on, speakers sensitive to this will be under pressure to use higher and higher indexes on the truth-predicate. Someone who wants to disagree with me will have to use an i still higher than mine, and so on.

Of course, followers of Burge might claim that which i we select for our interpretation of 'true' isn't up to the speaker; it's determined by other contextual elements, like principles of interpretative charity, or whatever. But in a way, this makes things worse. For then, to ensure that my intentions to say something about *all* of so-and-so's relevant utterances are satisfied, I can opt to say something long-winded, embedding semantic predicates within other semantic predicates. For example, if I assert: "For all s, if Jones says s, then either s isn't true, or it isn't true that s is true, or it isn't true that it's true that it's true that s is true," etc., then my statement simply can't be interpreted with a lower index on the truth predicate than there are disjuncts. But surely I add nothing to my thought by doing this.

Does Glanzberg's approach fare better than Burge's? He won't be subject to *exactly* the same problem, because he isn't in the business of recursively constructing T_is based on which T_js can go in its extension. But he'll have parallel problems. When we interpret your utterance and Jones', we'll have something like:

$$s_1 = Tiff(\forall s)(JonesAsserts(s) \rightarrow \neg(\exists_i p)[Exp(s,p) \& T(p)])$$

$$s_2 = \mathit{Tiff}(\forall s)(\mathit{SethAsserts}(s) \rightarrow (\exists_j p)[\mathit{Exp}(s,p)\&T(p)])$$

We must ask the same question we asked on Burge's construction: Whose existential quantifier ranges over more propositions? A principle of symmetry like Burge's might compel us to make $i = j$. But even so, when we reason about whether we should believe s_1, s_2, or neither, we're going to reason ourselves into loops much like those in the regular Liar inference. Roughly, your utterance will be true on (say) the even $\exists_i$ quantifiers, and Jones' will be true on the odd $\exists_i$ ones. This strikes me as unintuitive; if there *are* domains of truth conditions that render s_2 true, then you would not be content with your s_1; it would fail to be true in the sense in which you intended it. You would derive no comfort from the fact that your utterance expressed a true proposition on every even i on the metalinguistic quantification over propositions. The right thing to say, it seems, is that you weren't right at all, not that you were right in every other context on an infinite hierarchy.

It should go without saying that you can't get around this problem by quantifying out the contextual parameter, saying, 'There's *no* context in which Jones says something true!' For then the contradiction reappears: As long as Jones is clever enough to follow you in quantifying out the contextual parameter, we're back with the regular old contingent Liar, just a slightly more complicated one.

Again, maybe there's just no way for everyone to satisfy their intentions when they turn out, due to extremely unfavorable circumstances, to lead to paradox. But I would opt for a solution that didn't include this counterintuitive result.

2.9.2 Contingent paradoxes and epistemic attitudes

Leaving aside arms-races, contingently paradoxical sentences present another problem for contextualist solutions—this time one that affects Glanzberg's analysis more acutely than Burge's. This problem isn't that we can't satisfy our (perhaps perverse) intentions to talk about the truth-values of all the sentences that other people have uttered. Instead, it revolves around trying to make sense of our epistemic attitudes towards sentences that turn out to be (contingently) paradoxical.

Let's have a simple example. Disgusted by the weatherman's recent predictions, I wake up in the morning and declare, "The first sentence uttered by the weatherman on the 10pm news will be false." (Call this s_3.) I then remain silent the rest of the day. Unfortunately, news of my prediction travels wide, and the weatherman, a trickster, decides to open the 10pm news by saying s_4: "Everything that Ethan Jerzak said today was true."

The structure of these contingently paradoxical sentences is familiar: with classical inferences plus the standard truth rules, these two sentences allow us to derive a contradiction from no premises. The contextualist blocks the paradox by restricting the sentential truth-predicate to some T_i on a hierarchy. The question I want to ask is: How should we analyze my epistemic attitude toward s_3? As it turns out, it does not express a unique proposition; insofar as it expresses any proposition at all, it's one on an infinite series of merely seeming contradictions on an infinite contextual hierarchy. So consider the question: "Did I believe s_3?" Well, when I believe a sentence, I'm believing the *content* of that sentence—that is, the proposition expressed by that sentence, according to folks like Glanzberg. So on his analysis, this question is actually ill-formed: My sentence doesn't have a determinate content, so we cannot sensibly ask, "Did I believe *the* proposition expressed by s_3?"

But this is extremely counter-intuitive. Surely I *did* believe a determinate thought when I expressed s_3: I believed neither more nor less than that the weatherman would say something false. Glanzberg seems forced to fragment what is one belief into a series of beliefs, none of which expresses the thought that I wanted to express.

Of course, Glanzberg *could* say that s_3 expressed a determinate thought when I uttered it, but ceased to when the weatherman uttered his devious utterance. But I'm not sure whether this fares much better. When, exactly, do I go from believing a determinate proposition in believing s_3 to failing to? When the weatherman utters s_4? When I finally get around to watching the 10pm news on my Tivo at 11pm? Wherever the line is drawn, the result is weird. Our worries about contextualism as an *ad hoc* solution to the paradoxes are not yet quelled.

Therefore, Glanzberg owes us an analysis of these contingently paradoxical sentences and the kinds of epistemic attitudes that we can have toward them, before his proposal becomes compelling as a solution to the semantic paradoxes for natural language. A virtue of Burge's paper is that he deals with these contingently paradoxical sentences head-on; Glanzberg's proposal, as it is given in his 2001 and 2004, suffers from failing to do this. He discusses Liar-type reasoning only as it arises in the fairly artificially-constructed l. But we don't have a truth predicate in English so that we can assert weird, explicitly self-referential sentences like *l*. We have it precisely to talk about the truth-values of other utterances. These contingent paradoxes are therefore the true test-cases for theories of the English truth predicate, particularly as it interacts with the deductive system. Glanzberg should treat them as such.

2.10 Upshot

Introducing propositions allows the contextualist to respect the sense in which 'true' seems to have a single meaning: as applied to propositions, there's no indexicality in the word 'true'. Trouble is, the word 'true' in English can also be applied to utterances of sentences, and therefore the truth predicate that we actually reason with will be fragmented. In particular, contextualism makes T-in, taken literally, invalid; we only get the inference from s to $T_i(\ulcorner s \urcorner)$ for some contextually dependent i. This makes much of our reasoning about truth inferentially impotent: as we've seen, proving that a sentence is true$_i$ is a far cry from proving that a sentence is *true*.

Truth plays a more pervasive role than merely as applied to utterances of sentences. As we'll see in what follows, truth is intimately connected to our reasoning about knowledge and belief. Paradoxes structurally similar to the Liar can be formulated involving these notions, too, and it's less clear that a contextualist approach will be applicable in those cases. When these propositional attitudes are combined with self reference, I'll argue, contextualist solutions become less promising, and ones that abandon certain aspects of classical logic fare better.

Chapter 3

Non-Classical Knowledge

3.1 The Knower paradox

The Knower paradox purports to place surprising *a priori* limitations on what we can know. According to orthodoxy, it shows that we need to abandon one of three plausible and widely-held ideas: that knowledge is factive (FACT), that we can know that knowledge is factive (KFACT), and that we can use logical/mathematical reasoning to extend our knowledge via very weak single-premise closure principles (SPC).

In what follows, I argue that classical logic, not any of these epistemic principles, is the culprit. The plan: I draw out the structural similarities between the Knower and more familiar semantic paradoxes like the Liar. I extend one popular non-classical treatment of the Liar paradox to the Knower paradox, showing that all of these principles can be saved with conservative and philosophically motivated emendations to classical logic. Finally, I evaluate the resulting theory for plausibility, for knowledge of both mathematical and natural language claims. I consider and respond to two objections to my approach that arise in the mathematical context. The first objection is that the indirect nature of sentential reference via Gödel coding renders the formulation of (KFACT) needed to generate the paradox implausible. The second objection is that the way I construct my non-classical theory happens to take a stand on a complex question in the philosophy of mathematics known as Gödel's disjunction, and thus is attractive only to those antecedently disposed to that view. Finally I evaluate my proposal in the context of natural language knowledge attributions. Here, things are more promising: my proposal fits nicely with our intuitions about knowledge and reasoning, and the counterpart of the objection for mathematical knowledge concerning sentential reference does not get off the ground.

3.1.1 Background: a brief history of the Liar

It's natural to think of the Liar paradox as a foil for our intuitive concept of truth. The story goes something like this: Take any theory expressive enough to represent the primitive recursive functions. Such a theory has the resources to arithmetize its syntax, allowing for the expression of predicates of sentences in the language. Add a predicate, T, whose intended interpretation is that $T(\ulcorner\varphi\urcorner)$ holds just in case the sentence φ—the sentence, that is, whose Gödel number is $\ulcorner\varphi\urcorner$—is true. The Liar paradox shows that, whatever your theory of truth, it must not validate all instances of the axiom schema

$$T(\ulcorner\varphi\urcorner) \equiv \varphi, \qquad \text{(Convention-T)}$$

for then disaster follows. The diagonalization lemma guarantees us some sentence l such that it's provable that $l \equiv \neg T(\ulcorner l\urcorner)$. As is well known, triviality follows: contradictions, and hence all sentences, are theorems of the resulting theory.

Usually when adding an axiom to a consistent, well-motivated theory leads to contradiction, the natural response is to recommend against adding it. Though the T-schema seems intuitively like a non-negotiable part of any theory of truth, the Liar paradox shows that intuition to be unsalvagable. The task for theorists of truth is therefore to find the most respectable way to weaken Convention-T. The battle lines are drawn between the two directions of the bi-conditional:[1]

$$T(\ulcorner\varphi\urcorner) \to \varphi; \qquad \text{(T-out)}$$

$$\varphi \to T(\ulcorner\varphi\urcorner). \qquad \text{(T-in)}$$

According to this response, theorists of truth essentially have the job of deciding which of T-out and T-in to throw away. Gap theorists reject T-in,

[1] Sometimes I will speak of the *rules of inference* (rather than axioms) **T-out** and **T-in**, which are exactly what you would expect: A rule of inference from $T(\ulcorner p\urcorner)$ to p and vice-versa. Classically (or more generally in any logic with modus ponens and if-introduction) the unrestricted validity of these rules is equivalent to that of the schemas, but in the non-classical theories I'll be working with later on they come apart. When I mean the rule, I'll write it in bold; otherwise I mean the axiom.

and countenance the possibility of sentences with semantic value 1 which fail to be true. Glut theorists reject T-out, and countenance the converse possibility. Still other theories (like the revision theory) countenance some subtle mixture of these possibilities. But whatever the theory, fiddling with the T-schema seems initially unavoidable.

But only initially. Recent work on the semantic paradoxes (by, for example, Priest (1987), Maudlin (2004), and Field (2008)) has expanded the number of weapons in our arsenal for battling the semantic paradoxes, by weakening the background logic. The Liar allows us to derive contradictions when we add T-out and T-in to classical first-order logic. But classical logic was not handed down to us on stone tablets; why ought inferences like *reductio* be beyond rational reproach, while such obviously valid inferences as that from $T(\ulcorner p \urcorner)$ to p fall by the wayside? For no *a priori* reason, many have concluded. When we build theories for bits of language, we do so holistically, and no part of the theory, not even classical logic, is immune from rational revision. Thus, the initial story looks undermotivated as an inescapable consequence of the Liar. Perhaps we need not reject Convention-T and accept a revisionary and second-rate truth-predicate; perhaps we were simply wrong about the underlying logic in which to theorize about notions like truth.

Now, there's *prima facie* reason to expect a unified treatment for all paradoxes which involve Liar-like self-reference. It would be odd to take a truth-value-gap approach for the Liar, but a glut approach to Berry's paradox, and an emendation to classical logic for Curry's paradox. Whenever you've got a sentential predicate governed by intuitive but badly behaved axioms which jointly lead to contradiction, you should expect to block the derivation in more or less the same way that you blocked the Liar—absent, that is, some particular reason to think that the nature of the property represented by the predicate in question warrants a different approach. With this (very!) brief history of the Liar in view, let's consider a paradox with a quite different genealogy.

3.1.2 The Knower

The Knower paradox was discovered as a special case of the surprise exam paradox in Kaplan and Montague (1960).[2] Here's a standard way to present it.

[2] I'll be concerned here with the more purified forms of the paradox that have developed in the literature after Kaplan and Montague (1960). The way it arises out of the surprise

One of the things we talk about in English is knowledge. We epistemologists want to investigate its nature, and one of the ways to do that is to describe its structural properties. How should we proceed?

The most well-known way of formalizing our thought and talk about knowledge is with a sentential operator, as in modal logic. Syntactically, for every sentence φ, we add to the language a new sentence $K\varphi$, which says that φ is known. This approach has proved extremely productive in formalizing many aspects of knowledge. However, sentential operators are not expressive enough to capture all of the ways we express knowledge attributions. We can refer to the objects of knowledge indirectly, as in: "John didn't know the first thing he said yesterday." We can also quantify over the objects of knowledge, as in: "Anything that Sally knows, John knows too."

A standard sentential operator is not expressive enough to capture uses like these. In the first case, there is no particular sentence φ to put 'K' in front of; what follows 'K' is a noun phrase. While for some modeling purposes one may simply insert the actual sentence φ which John said, that won't have the same general truth conditions, for John might well have uttered a different sentence. Similarly for the quantified attributions: no finite number of sentences appended to the knowledge operator and then conjoined could have the same truth conditions as "Anything that Sally knows, John knows too," for Sally can always come to know something new. To capture the full expressive power of our thought and talk about knowledge, we need a knowledge *predicate*, that takes *terms* (which refer to sentences) as arguments.[3] That way we can express the whole range of knowledge attributions that English allows. So, let $K_\alpha(\ulcorner\varphi\urcorner)$ hold just in case rational agent α knows the sentence φ. We'll be concerned here only with a single agent, and so leave off the subscript.[4]

exam paradox is by interpreting the teacher's announcement that there will be a surprise exam sometime next week as implicitly self-referential, along the following lines: "Either there will be an exam sometime next week such that you won't have known it would occur on the morning it occurs, or else you can't know the very sentence I'm uttering right now." Subsequent discussions of the Knower paradox have focused on the isolated second disjunct.

[3] There are modal logics that allow for propositional quantification intended to capture uses like these. But as Stern (2014) points out, it is possible to force self-reference even with operators, by introducing fixed-point/diagonalization axioms. Thus attempts to avoid the paradoxes by restricting to operators don't get to the heart of the matter. Any non ad-hoc way of representing these kinds of knowledge attributions end up being paradox-prone, one way or another.

[4] For simplicity, we restrict here to sentences φ whose meaning does not depend on

What kinds of axioms and inference rules govern the behavior of this predicate? These seem like a good start:

$$K(\ulcorner \varphi \urcorner) \supset \varphi, \tag{FACT}$$

$$K[\ulcorner K(\ulcorner \varphi \urcorner) \supset \varphi \urcorner], \tag{KFACT}$$

$$\begin{array}{r|llr}
i & K(\ulcorner \varphi \urcorner) & & \\
j & \quad \underline{\mid\ \varphi\ } & & \\
\dots & \quad \mid\ \dots & \textit{(This subproof can use nothing derived from premises;} & \\
k & \quad \mid\ \psi & \textit{only the rules and axioms of theory X.)} & \text{(SPC)} \\
k+1 & K(\ulcorner \psi \urcorner) & \mathit{SPC}_X,\ i,\ j\text{-}k &
\end{array}$$

(FACT) is a central part of our concept of knowledge, to be abandoned only as an absolute last resort. Since all instances of (FACT) hold necessarily, there shouldn't be any trouble imagining an agent sufficiently schooled in epistemology who comes to know any particular instance of (FACT). Thus it shouldn't be impossible to imagine a rational agent described by (KFACT).

(SPC) requires some explanation. It formalizes the idea that α can extend her knowledge by impeccable reasoning by some sound theory X. Since we'll be comparing different theories for the purposes of this paper, we allow X to be variously permissive, with greater permissiveness resulting in stronger closure principles. For example, sometimes we might only require closure under propositional logic; we'd then let $X = PL$. Other times we might want the full resources of Peano arithmetic, and so let $X = PA$, or the full resources of PA with the addition of the factivity axiom (FACT): $X = PA + (FACT)$. The important thing is that under no circumstances are *premises*, or later steps which depend on premises, allowed in these closure subproofs. Only the rules and axioms from theory X are allowed. Our starting point will be to let $X = PA$, and that's what I'll refer to with unsubscripted uses of (SPC).

(SPC) is one of the weakest closure principles that you can cook up. Single-premise derivability in *PA* is a very strong relation, and though it's easy to imagine (SPC) failing because of an agent's limited computational

context. Otherwise we'd need a more complicated knowledge relation—something like, α knows φ as uttered at c.

powers, it's hard to imagine that this is anything other than a contingent computational limitation. Nothing in the nature of knowledge itself should prevent me from knowing a *PA*-consequence of any particular sentence which I know. It's worth noting that one popular kind of argument against closure, of the risk-aggregation lottery (Kyburg (1961)) or preface (Makinson (1965)) variety, does not affect single-premise closure.[5]

The problem with these three axioms is that they are (classically) inconsistent (Kaplan and Montague (1960)). The culprit is a self-referential sentence reminiscent of the Liar. The diagonalization theorem does its work, giving us a sentence g such that the following is PA-derivable:

$$g \equiv \neg K(\ulcorner g \urcorner) \qquad \text{(Knower sentence)}$$

Here's how to derive a contradiction using the Knower sentence using (FACT), (KFACT), and (SPC). Remember that $g \equiv \neg K(\ulcorner g \urcorner)$ is derivable from no premises in *PA*, and is therefore admissible as a derived axiom for the purposes of (SPC). Let $(FACT_g)$ be the instance of $(FACT)$ instantiated with the Knower sentence g, and similarly for $(KFACT_g)$; what follows is a proof of $\bot$, using (SPC).[6]

[5] What I'm calling single-premise closure is not the principle, sometimes identified in discussions of skepticism as such, that $[K(\ulcorner p \urcorner) \wedge K(\ulcorner p \supset q \urcorner)] \supset K(\ulcorner q \urcorner)$. This is strictly stronger than what is needed to get the Knower paradox going, at least on the standard idealizing assumption of logical omniscience.

[6] $\bot$ can stand for any arbitrary absurdity. For concreteness, let $\bot$ be true iff $0 = 1$.

1	$K(\ulcorner g \urcorner) \supset g$	(FACT_g)
2	$K(\ulcorner K(\ulcorner g \urcorner) \supset g \urcorner)$	(KFACT_g)
3	$\quad K(\ulcorner g \urcorner) \supset g$	
4	$\quad g \equiv \neg K(\ulcorner g \urcorner)$	Diagonalization
5	$\quad\quad K(\ulcorner g \urcorner)$	
6	$\quad\quad g$	Modus ponens, 3,5
7	$\quad\quad \neg K(\ulcorner g \urcorner)$	Modus ponens, 4, 6
8	$\quad \neg K(\ulcorner g \urcorner)$	Reductio, 5-7
9	$\quad g$	Modus ponens, 4, 8
10	$K(\ulcorner g \urcorner)$	*SPC*, 2, 3-9
11	g	Modus ponens, 1, 10
12	$g \equiv \neg K(\ulcorner g \urcorner)$	Diagonalization
13	$\neg K(\ulcorner g \urcorner)$	Modus ponens, 11, 12
14	$\bot$	$\bot$-intro, 10, 13

Thus $\bot$ is derivable in classical *PA* from $\{(\mathit{FACT}_g), (\mathit{KFACT}_g)\}$, and (SPC).

This result seems to doom principles (FACT)-(KFACT)-(SPC), showing that they cannot hold even of perfectly rational subjects. Either knowledge isn't always factive, or there are sentences for which you are barred in principle from knowing the corresponding instance of the factivity schema, or else knowledge is not closed under single-premise derivability in weak theories. Nobody seriously considers abandoning (FACT), so the literature on the Knower has been, so far, a back-and-forth between (KFACT) and (SPC). Maitzen (1998) argues for abandoning closure; Cross (2001) responds with a defense of closure and some reasons to give up (KFACT). Uzquiano (2004) rebutts Cross, suggesting that closure might be the culprit after all.[7]

The aim of this paper is to try out a different tack. It's clear that the

[7] An exception to this dialectic is Sainsbury (1995), who briefly discusses the possibility that the sentence, rather than the epistemic axioms, might be at fault. But he does not develop this thought very far there.

Knower is a paradox of the same basic kind as the Liar. We've got a sentential predicate governed by axioms that do not play well together when self-referential sentences are formulatable. Indeed, the knowledge axioms overlap with the problematic truth axioms: T-out has exactly the same form as (FACT), and (KFACT)-(SPC) together play basically the role of T-in, allowing us to derive sentences involving the problematic predicate from sentences without it. Given these structural similarities between the way the paradoxes get going, we should expect them both to stem from a common root. And a common root calls out for a common solution.

As discussed in the introduction, there is a wider array of options at our disposal for dealing with the Liar paradox than merely rejecting T-out or T-in. Whereas (KFACT)-(SPC) have marked the only disputed territory for solving the Knower, the Liar has inspired revisions to propositional logic itself. It would be a waste not to avail ourselves of these proposals in our search for a solution to the Knower; after all, if the best background logic in which to add a truth predicate blocks the Knower paradox even while (FACT)-(KFACT)-(SPC) or close analogues are all accepted, then we needn't spend time haggling over which to reject. The general lesson is simply that the literature on the Knower has overlooked live options for dealing with the paradox. I wish to put those options properly on the table.

I won't be concerned with all possible ways to use non-classical logics to defeat semantic paradoxes. Instead, I'll sketch one popular non-classical theory, and add the modal ingredients necessary to formulate a knowledge predicate in that theory. The paradox is indeed blocked; the philosophical question is whether the benefits of this approach outweigh its costs. I argue that in natural languages, the benefits do outweigh the costs, but that in the context of specifically mathematical knowledge, the question depends on unresolved issues about our ability in-principle abilities to know mathematical truths.

3.2 A solution, outlined

The theory of truth that I'll take as my starting point was developed in its essentials by Saul Kripke, and refined by Hartry Field. In this section, I'll give a fairly informal sketch of how his construction works. (Appendix A contains a detailed construction.) This theory has two elements: Kripke's truth predicate, and a novel conditional called the Field conditional. I show

how to construct a knowledge predicate out of a modal operator for necessity and the truth predicate we get from Field's theory, roughly along the lines Halbach and Welch [2009] take for a metaphysical necessity predicate and Caie (2012) for a belief predicate. I'll show that, in this construction, all of the ingredients that led to paradox in classical logic can peacefully coexist.

3.2.1 Kripke's truth predicate

Kripke's idea is to enrich an arbitrary classical model $\mathcal{M}$, which is completely silent about which sentences go into the extension of the truth predicate, into a souped up model $\mathcal{M}^+$ that matches $\mathcal{M}$ where the truth predicate isn't involved, but that gets the extension of the truth predicate right—that is, $\mathcal{M}^+$ assigns a sentence to the extension of the truth predicate just in case $\mathcal{M}^+$ assigns that sentence semantic value 1.[8]

The Liar paradox shows that such a fully deflationary truth predicate cannot exist in classical theories. Kripke therefore makes use of the strong Kleene connectives in a theory with three semantic values: 0, $\frac{1}{2}$, and 1. The definitions of the sentential connectives are as follows:

$$
\begin{aligned}
&[\![p]\!] = 1 \text{ if } v(p) = 1; 0 \text{ otherwise}\\
&[\![\phi \wedge \psi]\!] = \boldsymbol{min}\{[\![\phi]\!], [\![\psi]\!]\}\\
&[\![\phi \vee \psi]\!] = \boldsymbol{max}\{[\![\phi]\!], [\![\psi]\!]\}\\
&[\![\neg\phi]\!] = 1 - [\![\phi]\!]\\
&\varphi \supset \psi := \neg\varphi \vee \psi
\end{aligned}
$$

Note that these definitions give us exactly the classical Boolean connectives in the special case where all the semantic values of the constituent sentences are 0 or 1. This is important, for this feature allows this theory to vindicate particular instances of classical reasoning when there's no danger of paradoxical indeterminacy.

The logic K_3 is what we get by taking 1 as the designated semantic value: an argument is valid just in case all models which assign the premises

[8] It's important not to confuse "having semantic value 1" with "being true". The former is a metalinguistic notion, while the latter is our way of representing truth in the object language under scrutiny. One task for theorists of truth is to get these two notions to line up extensionally as much as possible. Also, note that the semantic theory for this non-classical language is given in a fully classical metalanguage. See Field (2008) for a discussion of how embarrassing this should be.

semantic value 1 also assign the conclusion semantic value 1.[9] As usual, valid sentences are those that follow from no premises.

A notable thing about K_3 is that there are no valid axiom schemas involving these connectives. That's because, for any of the connectives, if all of the constituents have semantic value $\frac{1}{2}$, the whole sentence has value $\frac{1}{2}$. So whereas in classical logic, you can swap axioms for inference rules, in K_3 there are tons of valid inference rules, but no axioms.

The classical rules that are not valid in K_3 are those which allow us to exit Fitchean sub-derivations: reductio and if-introduction. This is unsurprising, for it is these rules from which axioms can be derived from no premises in classical logic. Crucially for my treatment of the Knower paradox, reductio is not a valid rule of inference. If the sentence we suppose for contradiction entails $\perp$, that could be because it has value 0, or because it has value $\frac{1}{2}$; and in the latter case its negation will not have value 1, but will instead have value $\frac{1}{2}$. Similarly with (material) conditional introduction: we can suppose the Liar sentence, and validly derive the Liar sentence, but $T(\ulcorner l \urcorner) \supset T(\ulcorner l \urcorner)$ has semantic value $\frac{1}{2}$, not 1.

We build the souped up Kripkean model $\mathcal{M}^+$ in stages corresponding to temporary quasi-extensions for the truth predicate. The first quasi-extension T_0 is the empty set.[10] The first temporary Kripke model $\mathcal{M}^{+_0}$ uses T_0 as its extension for the truth predicate, matching $\mathcal{M}$ on the non-T-involving sentences, and assigning all sentences involving the truth predicate semantic value $\frac{1}{2}$. The next quasi-extension T_1 includes all of the sentences that the first temporary model $\mathcal{M}^{+_0}$ assigned value 1. $\mathcal{M}^{+_1}$ uses T_1 as its extension for the truth predicate, matching $\mathcal{M}^{+_0}$ on all non-T-involving sentences, but assigning semantic value 1, in addition, to $T(\ulcorner s \urcorner)$ for every s that $\mathcal{M}^{+_0}$ assigned semantic value 1, and semantic value 0 to $T(\ulcorner s \urcorner)$ for every s that $\mathcal{M}^{+_0}$ assigned semantic value 0.

You repeat this process throughout the ordinals, taking the union of all the previous quasi-extensions at limits. Eventually this process finds a

[9] Graham Priest's logic of paradox, LP, uses the same semantics, but lets both 1 and $\frac{1}{2}$ be designated values. See Field (2008), p. 79 for a complete presentation of K_3 and the relationship between it, LP, and classical logic. K_3 reduces to classical logic if you add $\phi \vee \neg\phi$ as an axiom schema.

[10] This gives us what's called the minimal fixed point. You can put tame self-referential sentences in here, if you want, to get larger fixed points. For example, no contradiction arises from supposing that the so-called 'truth-teller' sentence—"This sentence is true"–is in T_0.

fixed point, such that iterating the process another time does not add any sentences to the extension of the truth predicate.[11] The honor of Final Extension T^+ is given to this fixed point. The souped up model $\mathcal{M}^+$ uses this extension for the truth predicate: the semantic value of $T(\ulcorner\varphi\urcorner)$ in $\mathcal{M}^+$ is 1 if $\varphi \in T^+$; 0 if $\neg\varphi \in T^+$; and $\frac{1}{2}$ otherwise. The Liar sentence l is one sentence such that neither it nor its negation ends up in the fixed point; it therefore receives semantic value $\frac{1}{2}$. (The details of this construction are first worked out in Kripke (1975), and you can find a more succinct and cleaned-up presentation in Field (2008), §3.1. Appendix A also includes most of the details.)

This minimal fixed point construction, or some variant of it, constitutes the heart of many popular resolutions to the Liar paradox. It corresponds nicely to the idea that sentences like the Liar don't inherit their truth or falsity from the world in the right way. In order to tell whether l is true, we have to look at what l says; but what l says involves reference to l's truth-value. There's a vicious cycle of semantic dependence for sentences like this. The fixed-point construction of the truth predicate seems to explain this intuition: l never ends up in the minimal fixed point precisely because its truth isn't inherited in the right way from truth-free atomic sentences. Atomic sentences not involving semantic vocabulary are ultimately where language hooks up with the world, and the semantic value of the Liar never traces back to that of any atomic sentence. It is, so to speak, a frictionless spinning in the semantic void.

T-out and **T-in** are valid rules of inference on this construction. Why? **T-out** is easy: If $T(\ulcorner s\urcorner)$ has semantic value 1 at the fixed point level, that means that s was assigned to some quasi-extension of T, and we are careful to set up the quasi-extensions so that only sentences with semantic value 1 ever get admitted. To prove that **T-in** is valid, we just need to observe two things: First, we never lose any sentences as we go up the ladder of quasi-extensions; second, the final interpretation of T's extension is a fixed point of the inductive procedure. Thus if s has value 1 at the fixed point level, it will be assigned to the extension of T at the next quasi-extension. But the ultimate interpretation of T's extension was a fixed point, so we needn't wait until the next level: $T(\ulcorner s\urcorner)$ will already be in the fixed point. So

[11] The argument for this is a pretty simple brute-force one about size. This process continues through the ordinals, and you cannot keep adding sentences at every stage, because there are more ordinals than there are sentences of the language.

T-in is valid.

Nevertheless, not all instances of the T-schema, as formulated with the material conditional, hold. Why not? One such instance is:

$$T(\ulcorner l \urcorner) \equiv l$$

When l is the Liar sentence, this does not have semantic value 1; according to the strong Kleene connectives, it receives the same semantic value as its constituents, namely $\frac{1}{2}$.

3.2.2 Field's conditional

Kripke's procedure gets the extension of the truth predicate right, but all is not entirely well. The main shortcoming of his approach is that it's surprisingly hard to get a decent conditional out of the ingredients he gives us. The standard definition of the material conditional in terms of negation and disjunction is unacceptably weak in K_3, failing to validate even such trivialities as $\varphi \supset \varphi$. The obvious ways to add a stronger but still properly truth-functional conditional reinvite paradox (see Field (2008), Ch. 4). Halbach and Welch [2009] investigate how to add modality to Kripke's theory, but without a better conditional, the Knower paradox remains unsolved; there are *no* valid axiom schemas within the basic K_3 framework, and therefore (FACT) and (KFACT) couldn't be salvaged without doing something extra.

Field's insight—and the second main ingredient of his theory—is to give the conditional itself a revision-style semantics, greatly improving the conditional while still circumventing the possibility of unforeseen paradoxes. He formulates T-out and T-in in terms of this new, souped up conditional; I'll do the same with (FACT) and (KFACT) in a modal context, to generate a consistent, untyped knowledge predicate which validates these axioms unrestrictedly.

Again, the details of how to enrich Field's construction with modality are in Appendix A of this paper. It's nearly impossible to say what Field's conditional means informally—harder even than Kripke's truth predicate. The basic move is to splice together a fixed-point construction for the truth predicate, and a revision procedure for the conditional. You start out with the same base model $\mathcal{M}$, where the extension for the truth predicate is empty and every sentence whose main connective is $\rightarrow$ gets assigned semantic value $\frac{1}{2}$. Then, you build an entire Kripke fixed-point model to help with the truth predicate, yielding $\mathcal{M}^+$. That leaves the semantic values of sentences involving $\rightarrow$ untouched. After that, you're in a position to define the first temporary Field model $\mathcal{M}_1$, whose only work is to start fixing the right semantic values for $\rightarrow$: a sentence of the form $\varphi \rightarrow \psi$ gets semantic value 1 in $\mathcal{M}_1$ just in case, in $\mathcal{M}^+$, the semantic value of φ is less than or equal to that of ψ; otherwise it gets 0.

After you've finished building the first temporary Field model $\mathcal{M}_1$, you build an entirely new Kripke fixed-point model using $\mathcal{M}_1$ as the base, to get $\mathcal{M}_1^+$. Then you proceed as before to define the next temporary Field model $\mathcal{M}_2$: A formula of the form $\varphi \rightarrow \psi$ gets semantic value 1 in $\mathcal{M}_2$ just in case, in $\mathcal{M}_1^+$, the semantic value of φ is less than or equal to that of ψ; otherwise it gets 0. And so on. At limit stages, formulas of the form $\varphi \rightarrow \psi$ get semantic value 1 if the semantic value of φ is less than or equal to that of ψ at *all* previous stages; 0 if the semantic value of φ is greater than that of ψ at *all* previous stages; and $\frac{1}{2}$ if it fluctuates.

As before, you travel through the ordinals with this recipe in hand, splicing together temporary Kripke models (to chip away at T) with temporary Field models (to chip away at $\rightarrow$). A theorem proved in Field (2008), called the Fundamental Theorem, guarantees that this process eventually reaches some 'nice' stage η at which, according to $\mathcal{M}_\eta^+$, all the sentences have the 'right' semantic values, in a sense made precise in Appendix A. Models like this form the heart of Field's theory of truth; the semantic consequence relation for the theory is defined relative to the class of models with this property.

This theory does better than the standard non-classical Kripke theory. While the basic non-classical Kripkean framework can validate the rules **T-out** and **T-in** but not the axioms T-out and T-in, Field's theory validates all four, formulated with this special conditional. This makes it a good place to start when trying to validate other classically dangerous axioms involving a conditional, like (FACT) and (KFACT). That should be enough of a sketch of Field's theory of truth to get along with; here, our primary concern is not with the technical details but rather with its application to the Knower paradox.

3.2.3 The non-classical Knower

If we want to investigate non-classical solutions to the Knower along the lines of Kripke/Field style solutions to the Liar, we have two options. On the one hand, we could keep knowledge as a basic predicate of sentences, conjuring up a new fixed-point style construction so that semantically ungrounded sentences like g never end up in its extension. Or we could be more parsimonious, and use a combination of a standard knowledge-operator with a Kripke style truth predicate to get a knowledge predicate, using the already developed fixed point construction.

In what follows, I use a truth predicate in conjunction with an epistemic operator to define a knowledge predicate. While it might be somewhat unnatural to interpret the knowledge predicate by combining operators and truth, splitting up the predicate into an epistemic operator and a truth predicate has a practical advantage: epistemic operators and truth predicates are already well-explored phenomena. Combining them is a more familiar starting point than scratch.[12]

The most natural way to do this is to say that you know a sentence s just in case you know that it is true:[13]

$$K(\ulcorner s \urcorner) := \Box T(\ulcorner s \urcorner)$$

The generalization of Field's construction to include modality is straightforward. The only syntax that we need to add is '$\Box$'. Semantically, we start off with slightly fancier base models $\mathcal{M}$. We'll need each model to come with a non-empty set W of worlds, and a relation R between worlds. Since we eventually want $\Box$ to build us a knowledge predicate, and knowledge is factive, we'll require that the relation R be reflexive. (That is all that we'll require, which means that the base models are governed by the modal logic **T**.)[14] An interpretation function v sends predicate-world pairs to subsets of

[12] This issue is discussed at length in Halbach and Welch [2009] for metaphysical necessity, where it is shown that the most natural way to construct a necessity predicate from scratch is equivalent, via a translation function, to combining an operator with truth in the ways described below. So if, in your heart of hearts, you prefer a primitive knowledge predicate instead of a combination, you could take what follows as a mere consistency proof of (FACT)-(KFACT)-(SPC), without making any assumptions about whether combining operators with truth is really the right way to go on some more fundamental level.

[13] The other, slightly less natural translation would be:

$$K(\ulcorner s \urcorner) := T(\ulcorner \Box s \urcorner)$$

In what follows, nothing important hangs on this difference. Indeed, in the theory developed in Appendix A, these two formulations are equivalent. So in what follows I'll stick with the first, more natural construction.

[14] I do not assume that other epistemic principles, like for example $K\varphi \rightarrow KK\varphi$, shouldn't be part of our epistemic theory. I exclude principles like this here because they don't play any role in the Knower paradox.

the domain, and names to elements of the domain. Finally, μ is a standard variable assignment function which takes variables to elements of the domain. Thus formulas are assigned semantic values relative to a model $\mathcal{M}$, world w, and variable assignment function μ. The semantic clause for $\Box$ is exactly what you would expect, in the generalized K_3 context:

$[\![\Box\varphi]\!]^{\mathcal{M},w,\mu} = \min\{[\![\varphi]\!]^{\mathcal{M},v,\mu} : wRv\}$

That's basically all you need to do to enrich Field's language with modality. The procedure for building a Kripke fixed-point (for the truth predicate) and then a Field fixed-point (for the conditional) is basically untouched—the only difference is that the extension of the truth predicate is relative to worlds, since different formulas are true in different worlds. Similarly the fixed-point construction for the conditional fixes semantic values for $\rightarrow$ world-by-world. An inference is **valid** on this fancier construction just in case every world of every souped up Field model that assigns semantic value 1 to the premises also assigns semantic value 1 to the conclusion. A sentence is valid if the inference from $\varnothing$ to it is valid. The logic of this consequence relation includes (at least) all the rules of K_3, plus all of the axioms that govern the Field conditional. It also includes some axioms that mix the conditional, the truth predicate, and the box operator; some of these are exactly the axioms of interest in this paper, and are discussed below. Call the deductive system for this semantic consequence relation K$_3$FT ('K$_3$' for K$_3$, 'F' for Field, and 'T' because this logic extends the modal system **T** in bivalent settings).

The upshot: The following three propositions are the key results, at least as far as the Knower paradox is concerned, for this construction:

Proposition 3.2.1. All instances of the following schema are valid:

$$\Box T(\ulcorner\varphi\urcorner) \rightarrow \varphi.$$

Proof. See Appendix A. □

Proposition 3.2.2. All instances of the following schema are valid:

$$\Box T(\ulcorner\Box T(\ulcorner\varphi\urcorner) \rightarrow \varphi\urcorner).$$

Proof. See Appendix A. □

Proposition 3.2.3. The following rule is valid:

$$
\begin{array}{|l l}
\Box T(\ulcorner\varphi\urcorner) & \\
\begin{array}{|l} \varphi \\ \hline \dots \\ \psi \end{array} & \text{(Only axioms and rules of K}_3\text{FT.)} \\
\Box T(\ulcorner\psi\urcorner) & \text{Closure}_{K_3FT}.
\end{array}
$$

Proof. See Appendix A. □

Note that, on our interpretation of the knowledge predicate as a combination of the box operator and the truth predicate, these amount almost exactly to the ingredients that led to paradox in classical logic. Almost, because we use a different X here for SPC_X. I can't *exactly* validate full SPC_{PA}, because *PA* as such is formulated within a classical background logic which permits unrestricted reductio and if-introduction, even when the vocabulary involved in such proofs extends beyond pure arithmetic. However, the theory preserves the spirit of, and fundamental intuition behind, SPC_{PA}, which is that correct mathematical reasoning is a sure means to extend mathematical knowledge. All sentences formulated *purely* in the language of *PA* (we stipulate) have semantic value 1 or 0, and, as noted above, in such contexts K_3 reduces to classical logic. So non-classicality is only in play for sentences—dangerous, possibly ungrounded sentences—involving □ and T. Thus any use of SPC_{PA} to expand knowledge of purely arithmetic truths goes through on SPC_{K_3FT}; the only instances disallowed are ones which illicitly sneak a reductio or if-introduction subproof involving modal or semantic vocabulary into an *SPC* subproof—moves which aren't at all motivated by the intuitions behind SPC_{PA}. Such moves simply don't preserve semantic value 1 on the theory on offer; where there's semantic or modal vocabulary involved, there's the possibility of Liar-like ungroundedness, in which case rules like reductio aren't sound.

So, interpreted in this way, Proposition 4.1 is the factivity schema, 4.2 is the knowledge-of-factivity schema, and SPC_{K_3FT} is the closure rule. Those were the three ingredients that went into deriving the Knower paradox; *prima facie* one of these had to be given up to avoid contradiction. Does the Knower paradox, then, show that Field's construction (plus modality) doesn't work, since it validates all three of these principles?

No. $\bot$ is not a consequence of (FACT), (KFACT), and (SPC). I'll give the model-theoretic justification for this in a moment, first let's see exactly where the above proof fails. It fails at the reductio step:

5	$\underline{K(\ulcorner g \urcorner)}$		
6	g	Modus ponens, 3,5	
7	$\neg K(\ulcorner g \urcorner)$	Modus ponens, 4, 6	
8	$\neg K(\ulcorner g \urcorner)$	Reductio, 5-7	(not valid in K_3!)

In the closure subproof, we used the fact that g entails $\neg g$ to derive $\neg g$. But reductio is not a valid rule of inference on this construction. There are sentences (like the Liar and Knower) that entail contradictions, but have semantic value $\frac{1}{2}$ in all worlds of all models. The negation of a sentence with semantic value $\frac{1}{2}$ likewise has semantic value $\frac{1}{2}$. So this particular proof is fallacious.

But perhaps we are not being clever enough. How do we know that there's not some new proof of $\bot$ in K$_3$FT using the Knower sentence? Here is how we know: We defined (see Appendix A for details) an explicit model-theoretic construction, and the proof system K$_3$FT is sound on it. Since $\bot$ gets assigned semantic value 1 in none these models, no proof system that is sound on this class of models can possibly contain a proof of $\bot$. Of course, it takes some doing to show that the proof system *is* sound on this class of models; a small part of this work—the part crucial for the Knower paradox—is undertaken in the proofs of Propositions 4.1, 4.2, and 4.3 in Appendix A. Proving soundness for the more bread-and-butter rules and axioms is straightforward. The point is: we've built explicit models, and the proof system is sound on them. Thus there is no proof of $\bot$ from (FACT), (KFACT), and (SPC_{K_3FT}) in K$_3$FT, despite the fact that it permits Liar and Knower sentences.

The take-home point is this: The assumption, hitherto unquestioned in the Knower literature, that either (FACT), (KFACT), or (SPC) must be rejected (or else sentences like the Knower banished from the language) in order to steer clear of paradox is false. There is another option on the table: Use the already-developed Kripke construction for the truth predicate, along with Field's conditional, and formulate a knowledge predicate in terms of a box operator and the truth predicate. If you set things up this way, you will validate the natural reformulations of (FACT)–(KFACT)–(SPC), without ever being able to prove $\perp$ from them.

3.2.4 Evaluating this approach

Is this a *resolution* of the Knower paradox? It's tempting to say that it is. After all, I've shown that there is a perfectly consistent epistemic logic according to which knowledge, represented as an unrestricted, untyped predicate, is factive, it's known that knowledge is factive, and single-premise closure holds. Those are the three principles that the Knower paradox was supposed to have shown jointly inconsistent. What more could one expect of a resolution to a paradox?

This comes, of course, at a cost: We had to retreat to a background logic less powerful than classical logic, and use a special conditional to formulate the epistemic axioms. Is this a cost worth bearing? As in the case of the straightforward Liar, this approach should be evaluated holistically. What matters is which package, considered as a whole, provides a better model of our epistemic and inferential practices. Should we accept surprising limits on what we can know? Or should we weaken the rules of reductio and if-introduction, holding that they are sometimes unsound when the sentences they involve are semantically ungrounded? Discussions of this general question exist in Maudlin (2004), Field (2008), Bacon (2013), and (from the paraconsistent angle) in Priest (1987). Instead of rehearsing the general discussion about the payoffs of classical vs. non-classical logics for dealing with semantic paradoxes, I'll focus here on the novel issues that arise by formulating the paradoxes in terms of intentional notions like knowledge, as opposed to extensional notions like truth.

I argue for a somewhat surprising conclusion: that, unlike in the case of the extensional paradoxes of truth, intensional paradoxes like the Knower raise novel issues concerning the mechanisms of sentential reference. Different philosophical issues arise whether we formulate the paradox using direct sentential reference in natural languages, or using indirect Gödel coding in formal arithmetic/syntactic theories. These differences affect strength of the case for the non-classical approach developed above. In the natural language context, the epistemic principles are harder to abandon, and classical logic easier to abandon, than in the mathematical context. So ultimately, I argue, my construction has the definite edge in natural language settings. In mathematical settings, the jury remains out, and depends on complex issues in the philosophy of mathematics.

3.3 Mathematical knowledge

Standardly, semantic paradoxes like the Liar and the Knower arise against the background of syntactic/arithmetic theories. The object language under study has arithmetic vocabulary governed by mathematical axioms, and it is to this background theory that we add a predicate representing knowledge. The epistemic predicate introduced in such a context most naturally represents knowledge with specifically mathematical content.

In this section, I'll evaluate my solution from the mathematical perspective. First I argue that the intensionality of sentential reference via Gödel coding gives reason to doubt (FACT) is really a conceptual truth, thus calling (KFACT) into question for actual mathematical knowledge. I then ask whether the case for (KFACT) is stronger for *ideal* mathematical knowledge, or “absolute knowability" (as Koellner (2016) calls it). There, I show that my construction takes a stand on a complex issue in the philosophy of mathematics known as Gödel's disjunction. Thus I don't have a definitive case for my construction in the mathematical context. For such a case, I turn to natural languages, where the complications of indirect sentential reference via Gödel coding can be avoided.

3.3.1 Is (KFACT) plausible for mathematical knowledge?

In the literature on the Knower paradox, (KFACT) is generally motivated in the following way. The factivity of knowledge seems like a fundamental part of our concept of knowledge. It's a conceptual necessity if ever there was one. Thus, we should be able to come to know any instance of (FACT) simply by reflecting on this conceptual necessity. For surely, the thought goes, if there's *anything* we ought to be able to know, it's conceptual truths! Thus Uzquiano:

> A little reflection on (FACT$_g$) should convince any sophisticated epistemic subject that it is true. Hence we have every reason to think that (KFACT$_g$) is true. (Uzquiano (2004), my labeling of sentences)

However, in this section, I want to present some reasons for doubting that the formulation of (FACT) needed to generate the Knower paradox really does express a simple conceptual truth about knowledge.[15] This correspondingly gives us reason to doubt the plausibility of (KFACT), at least when 'K' represents the mathematical knowledge of actual agents. The crux of my argument is the very indirect kind of sentential reference that results from Gödel numbering, and the implications of this indirect sentential reference for knowledge attributions.

Recall that the first step in motivating the Knower paradox is to switch from a knowledge operator to a more expressive knowledge predicate. A knowledge predicate is needed to model straightforward indirect/quantified attributions of knowledge, but it also leads to paradoxical sentences like the Knower sentence. Now, when we made this move, we went from straightforwardly *using* sentences in knowledge attributions, to *referring* to them via some sort of referential mechanism. And in the context of mathematics, the kind of reference involved is very indirect. Mathematical languages like that of *PA* don't talk about their sentences directly—*PA* itself just talks about numbers. Instead, sentential reference is achieved by the indirect means of Gödel coding, which systematically assigns sentences to numbers.

The kind of sentential reference that results is by nature highly indirect, as as Halbach and Visser [2014a] and [2014b] have illustrated at length. They've examined three "sources of intensionality" in this kind of sentential reference, of which the one most relevant here is the first: the choice of coding.[16] Which numbers denote which sentences depends on the choice of coding scheme. And this choice is a highly arbitrary and contingent matter.

Halbach and Visser are interested in the implications of these sources of intensionality only for purely mathematical properties like truth and theory-relative provability. What's the upshot for the Knower paradox, where the

[15] Cross (2001) and [2004] also doubt that (FACT) is a conceptual necessity, but he does so only as the result of a modus tollens argument, and doesn't give an explanation for why it fails to be a conceptual necessity in terms of the intensionality of Gödel numbering.

[16] The other sources of intensionality are these. Source 2: Even relative to a fixed coding, there are multiple ways to formalize what it is for an arithmetic formula $\varphi(x)$ to express a property P of sentences. Source 3: Even fixing both the coding and the formalization of property-expression, there are multiple ways to formalize the notion of a self-referential sentence, depending on the particulars of the diagonalization proof. These other sources also matter for knowledge, for they make a difference to truth: relative to different particular diagonalization proofs, "the" sentence that asserts its own provability may be either provable, contingent, or refutable.

intended interpretation of 'K' is "is known"? The key point here is that every source of intensionality is a source of potential ignorance. It seems intuitively very hard to deny that we can know that knowledge is factive. However, when a knowledge predicate applies to sentences indirectly via Gödel coding, it's not clear that this general thought is faithfully expressed by all the instances of $K(\ulcorner K(\ulcorner\varphi\urcorner) \rightarrow \varphi\urcorner)$. Intentional notions like knowledge and belief are generally highly sensitive to the *way* in which terms refer. I may fail to know that Superman is Clark Kent, even if Superman and Clark Kent are necessarily identical. By the same token, I may fail to know that a sentential term $\ulcorner\varphi\urcorner$ denotes φ, even if $\ulcorner\cdot\urcorner$ is a computable coding function. And plausibly, I need to know what the term $\ulcorner\varphi\urcorner$ denotes, if I'm to know principles of semantic ascent and decent involving it.

To illustrate, consider (KFACT) in the context of straightforward sentential reference by definite description in natural languages. Let "Mary came to the party last night" be the first sentence uttered by John yesterday. Then an instance of (KFACT), using the non-rigid term "the first sentence uttered by John yesterday", would be: "I know that, if I know the first sentence uttered by John yesterday, then Mary came to the party last night". This can easily be false, if I fail to know what sentence was the first sentence uttered by John yesterday.

In this context, $\ulcorner g\urcorner$ is a metalinguistic name; corner quotes do not appear in the object language of *PA* plus the knowledge predicate ($\mathcal{L}_{PA+K}$). Literally, g is just some formula in $\mathcal{L}_{PA+K}$, and $\ulcorner g\urcorner$ is some numeral (which numeral it is depends on the coding). So in the actual object language, (FACT) will really look something like:

$$K(\overline{587}) \rightarrow \exists x(x = 0) \tag{3.1}$$

Well, probably with a much larger number on the left and a much more complicated mathematical formula on the right. But that will be its form: if you know the sentence coded $\overline{587}$, then $\langle$some formula in $\mathcal{L}_{PA+K}\rangle$. Now, this will certainly be *true*, because of how we set up the coding. We're interpreting K from a perspective external to the agent herself, giving a sideways-on account of the sentences which she knows. But, crucially, it says what it does partly because of facts about how we chose to set up the coding.

(KFACT), then, has the form:

$$K(\overline{3832}) \tag{3.2}$$

where $\overline{3832}$ decodes (relative to the coding scheme we chose) to (3.1). So when we flesh out (KFACT) literally in $\mathcal{L}_{PA+K}$, it no longer looks like the truism it misleadingly seemed like with metalinguistic corner quotes. (KFACT) was supposed to say that you know that knowledge is factive. That is, it's supposed to say that you know that whatever you know is true. This is plausibly a necessary truth that we should put into our theory of knowledge. But to know (FACT) as it's actually given in $\mathcal{L}_{PA+K}$, you'd have to know something else; you'd have to know *that* $\overline{587}$ decodes to $\exists x(x = 0)$. If there are live epistemic possibilities for you that a different coding system was used, you wouldn't know this, because you aren't sure which numerals decode to which sentences. For all you know, $\overline{587}$ could have denoted $\neg 0 = 0$! Thus to know (FACT) as formalized, you have to do more than simply reflect on the concept of knowledge; you yourself have to walk through the (possibly very long) proof in *PA* that, relative to the chosen coding scheme,

$$PA \vdash K(\overline{587}) \to \exists x(x = 0). \tag{3.3}$$

Now, it's not clear that there's any *in principle* bar to your doing this; after all, the coding function is computable. But the point is, (FACT) isn't the merely conceptual truth about knowledge it once looked like. There are substantial mathematical facts about how the sentences of *PA* are being referred to via $\ulcorner \cdot \urcorner$ involved.[17] If there's any plausibility to (KFACT), then, it isn't when 'K' is interpreted as actual mathematical knowledge of any particular agents, who don't concern themselves with reckoning out the exact denotations of long numerals relative to particular coding schemes. Instead, it's only plausible when 'K' is interpreted as some kind of highly idealized notion of in-principle knowability for an ideal mathematical reasoner, for whom such patience is not wanting. It's to this conception of 'K' that I now turn.

[17] Note the contrast to an operator representation of knowledge. The modal principle

$$K(K\varphi \to \varphi) \tag{3.4}$$

does not suffer from the same problem. Here φ is being used twice, not used once and mentioned once (as in (KFACT)). So (3.4) asserts no particular metalinguistic knowledge, while (KFACT) presupposes that the agent knows metalinguistic facts about what sentence the term $\ulcorner g \urcorner$ decodes to.

3.3.2 Is (KFACT) plausible for the ideal/absolute knowability of mathematical truths?

As we saw in the previous section, it takes more than mere reflection on the nature of knowledge to know instances of (FACT). It also requires some mathematical reasoning—in particular, actually walking through proofs of things like (3.3), which may involve a substantial amount of mathematical work. Thus ascribing (KFACT) to actual agents is highly implausible; there's no guarantee that such agents will have performed the specific proofs corresponding to (3.3) under the relevant coding scheme. However, since the coding function is computable, there's hope that this limitation is merely a contingent fact about actual agents. Maybe when 'K' represents mathematical knowability *in principle*, (KFACT) becomes more plausible. Thus the question: how far do the limits of our in-principle mathematical knowledge extend? Do we have reason to think that they extend far enough to include the instances of (KFACT) necessary to generate the Knower paradox?

The question about the in-principle limits of our mathematical knowledge goes back at least to Gödel's philosophical reflections on his incompleteness theorems. Koellner (2016) provides the the most extensive contemporary exposition, and I partly follow his dialectic here. The motivating thought is this. The incompleteness theorems show that mathematical theories strong enough to encode their syntax will inevitably have certain "blind spots". *PA*, for example, cannot prove that it is consistent, on pain of inconsistency. However, many have thought that *we* can know that *PA* is consistent, for we have an intended model of it in mind (zero, followed by its successor, followed by *its* successor...). We know that theories with non-trivial models aren't inconsistent. So we can know *Con*(*PA*), even if *PA* itself can't prove it. Our in-principle mathematical knowledge extends, in a sense, beyond what *PA* can prove.

Thus the question: Can in-principle mathematical knowability be captured by *any* finitely axiomatizable theory, perhaps one stronger than *PA*? Let **F** be an arbitrary finitely axiomatizable theory containing *PA* in $\mathcal{L}_{PA+K}$, let **K** be the set of sentences in this language ideally knowable by us, and let **T** be the set of true sentences in this language. We restrict to theories **F** which are themselves absolutely knowable and assume that what's absolutely knowable is true, which automatically yields the following relationship between these sets:

$$\mathbf{F} \subseteq \mathbf{K} \subseteq \mathbf{T} \tag{3.5}$$

Now, we know from the incompleteness theorems that $\mathbf{F} \subsetneq \mathbf{T}$—there are true sentences of $\mathcal{L}_{PA+K}$ that aren't contained in $\mathbf{F}$. Thus we know that at least one of these inclusions is improper: either $\mathbf{F} \subsetneq \mathbf{K}$, or $\mathbf{K} \subsetneq \mathbf{T}$. This is the core of Gödel's disjunction. If $\mathbf{F} \subsetneq \mathbf{K}$ for every $\mathbf{F}$, this means that our in-principle ability to know mathematical facts can't be captured by any finitely axiomatizable theory. If there is an $\mathbf{F}$ which coincides with $\mathbf{K}$, then $\mathbf{K} \subsetneq \mathbf{T}$. In this case, mathematical truth transcends our ability to come to know it—which entails a version of mathematical realism, on which at least some mathematical truths hold independently of us and our ability to prove them. Thus Gödel describes the disjunction in the following way:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised by a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the power of any finite machine, or else there exist absolutely unsolvable diophantine problems. (Gödel (1995))

Koellner (2016) shows how to formalize and then prove this disjunction given very plausible assumptions. Gödel himself believed the first disjunct, as reported in conversation:

> If one could clear up the intensional paradoxes somehow, one would get a clear proof that mind is not [a] machine. (Reported in Wang (1997), p. 187)

The Knower paradox is exactly kind of intensional paradox which Gödel thought stood in the way of proving the first disjunct. Penrose (1994) has famously attempted to provide a proof of the first disjunct, although his proof has been widely criticized (for example by Chalmers (1995) and Shapiro (2003)), and Koellner (2016) gives convincing reasons for doubting that such a proof is possible. In particular, relative to the same formal theories in which one can prove the disjunction itself, each disjunct is provably independent.

I by no means attempt to settle this issue here. Instead, I'll be content to situate my proposal with respect to the disjunction. I'll show that my proposal is much friendlier to the $\mathbf{F} \subsetneq \mathbf{K}$ disjunct. For those who insist on recursively enumerable in-principle knowability, then, I turn to natural languages. There, I argue, the Knower paradox arises in much more straightforward ways that bypass the subtle issues that arise in the mathematical setting.

The way I've justified the (non-classical!) consistency of (FACT), (KFACT), and (SPC) is to construct models (see Appendix) which validate them without validating $\bot$. Now, as sketched above, these models use a basic possible-worlds framework as a starting point for modeling epistemic agents. As is well-known, such a framework yields a *highly* idealized conception of epistemic agents. Since mathematical truths are necessary, they are true in every world of every model. Thus these models are ones on which absolute knowability includes every mathematical truth, and thus the set of ideally knowable sentences that results is definitely not recursively enumerable. Those who follow Gödel in being comfortable with this disjunct won't have a problem with this, and thus, for such people, my non-classical construction has definite value.

What about those who are committed to the other disjunct ($\mathbf{K} \subsetneq \mathbf{T}$), holding a conception of absolute knowability that *is* recursively enumerable?[18] Of course, while the particular possible-worlds models I've constructed happen to validate the $\mathbf{F} \subsetneq \mathbf{K}$ disjunct, this doesn't show that this disjunct *follows* (non-classically) from (FACT), (KFACT), and (SPC). It might be possible to construct non-classical models validating the other disjunct—maybe ones using impossible worlds (a standard move to deal with the problem of mathematical omniscience in a possible worlds framework). However, I'll conclude this section by offering an independent reason for doubting that a construction validating (KFACT) and the $\mathbf{K} \subsetneq \mathbf{T}$ exists. I'll show why those committed to this disjunct have reason to be independently suspicious of (KFACT).

Let us suppose that we disagree with Gödel, and hold that there must be some finitely axiomatizable theory $\mathbf{F}$ such that $\mathbf{F} = \mathbf{K}$, and assume that every instance of (KFACT) holds. It seems that there is a tension with Gödel's second incompleteness theorem lurking. Gödel's second incompleteness theorem shows that no recursively enumerable mathematical theory containing *PA* can prove its own consistency. Thus,

$$\mathbf{F} \nvdash \neg \mathit{Prov}_{\mathbf{F}}(\ulcorner 0{=}1 \urcorner) \tag{3.6}$$

however, as an instance of (KFACT),

$$K(\ulcorner K(\ulcorner 0 = 1 \urcorner) \supset 0 = 1 \urcorner) \tag{3.7}$$

[18] Uzquiano (2004) explicitly commits himself to this, but he does not engage there with the literature on Gödel's disjunction.

thus, because $\mathbf{F} = \mathbf{K}$,

$$\mathbf{F} \vdash K(\ulcorner 0 = 1 \urcorner) \supset 0 = 1. \tag{3.8}$$

But we can't have

$$\mathbf{F} \vdash \mathrm{Prov}_{\mathbf{F}}(\ulcorner 0 = 1 \urcorner) \supset 0 = 1, \tag{3.9}$$

because that would contradict (3.6) by the fact that $\mathbf{F} \vdash \neg 0 = 1$ and a simple application of modus tollens. However this doesn't quite refute (KFACT), because I may not know *that* **F** aligns with what I know: We'd get an actual inconsistency by adding $K(\ulcorner \mathrm{Prov}_{\mathbf{F}}(\ulcorner \varphi \urcorner) \supset K(\ulcorner \varphi \urcorner)\urcorner)$, which combined with (3.8) would entail (3.9).[19] Nonetheless, this might cause one to cast suspicion on (KFACT), for it denies that our in-principle knowledge has the kind of "blind spot" we know *all* recursively enumerable theories to have. So once the Knower paradox appears, (KFACT) immediately looks like the natural culprit, not classical logic. After all, if no Turing machine rigged up to compute the validities of some mathematical theory can spit out the sentence which says that that theory is consistent, why should I expect to be able to know, not only that my knowledge is consistent, but that it is factive?

This only scratches the surface of the subtleties involved in Gödel's disjunction. However, I won't belabor this point further. Instead, I'll switch gears, and argue that a much simpler and more definitive case for my non-classical approach can be made in the context of knowledge attributions in natural languages—one that sidesteps these subtleties.

3.4 The solution, evaluated: natural language

The philosophical landscape looks quite different in the context of natural language. Natural languages like English don't require anything fancy like a theory of syntax or Gödel coding to talk about their own sentences.

[19] As Koellner (2016) shows, this marks a difference between knowing that *some* Turing machine represents the set of sentences knowable by you, and its being knowable *which* Turing machine that is (relative to some enumeration of them). Koellner shows that it is provably inconsistent for someone to know *which* Turing machine represents the set of sentences knowable by him, but not inconsistent for someone to know that *some* Turing machine does.

While mathematical theories can *simulate* self-reference via coding and diagonalization, natural languages can do it straightforwardly and directly with demonstratives ('this very sentence...'), definite descriptions ('the first sentence uttered by John last Tuesday...'), and straightforward quantificational devices ('most sentences John said at the party...'). This, I'll argue in this section, is enough to sidestep the concerns of indirect sentential reference and ideal mathematical knowledge, and to make a case for my non-classical solution that doesn't depend on taking any particular stand on Gödel's disjunction.

Consider a straightforward, empirical version of the Knower paradox:

> (*) You don't know the first starred sentence on the page you're reading.

Do you know this sentence, or don't you? You seem to know the following things:

(A) If you know (*), then (*) is true. (Knowledge is factive)

(B) (*) is true if and only if you don't know (*). (Plain fact about what (*) says)

This enough to cause disaster in classical logic, by natural language reasoning analogous to the formal proof of $\bot$. (A) and (B) straightforwardly entail in classical logic (using reductio) that you don't know (*). Thus if you know the conjunction of (A) and (B), (SPC) says that can use this reasoning to come to *know* that you don't know (*). But this entails, together with your knowledge of (B), that you can also come to know that (*) is true. But then you are in the terrible position of not knowing (*), but knowing that (*) is true.

It might be tempting to call this impossible and deem the result a paradox, perhaps by invoking some principle like:

$$K(T(\ulcorner\varphi\urcorner)) \leftrightarrow K(\varphi) \tag{3.10}$$

There's some intuitive pull to the idea that to know that a sentence is true just is to know what it says. But this would be a mistake. It's not impossible to know that a sentence is true without knowing that sentence. This principle fails when you don't know what the sentence in question says. Say that I overhear a trusted scientist talking about some subject I don't understand. Maybe she says something like:

> (Q) Quarks are spin-1/2 particles.

If I trust her, I can plausibly come to know that (Q) is a true sentence. But I don't have any idea what (Q) means. So it would be wrong to attribute to me knowledge *that* quarks are spin-1/2 particles. To know that presupposes that I have some general idea what quarks are, and what it means for something to be a spin-1/2 particle. When she utters (Q), I don't have a clear enough sense of what possibilities are being ruled out to be said to know the content of what she's said, even if I'm certain that it's true, whatever it means.

This is even clearer when truth is attributed to an utterance by referring to it in non-quotational ways. Say that I have a scrupulous friend John who is well known never to say untrue things, and to say at least one thing every day. I can be generally aware of his trustworthiness and loquacity, and so know:

> (F) The first thing that John said this morning is true (whatever that was).

Now, say that the first thing John said this morning after waking up was:

> (G) Germany invaded Poland on 1 September 1939.

I may not have any idea exactly when Germany invaded Poland, so I may not know (G). But this doesn't stand in the way of my knowing (F), even though (G) *was* the first thing that John said this morning. I don't have to know exactly what John said in order to know that, whatever it was, it was true. Therefore, where $S(x)$ is the predicate abbreviating "x was the first sentence John said this morning", we have a definitely true instance of:

$$K(\ulcorner T(\imath x S(x))\urcorner) \wedge \neg K(\imath x S(x)) \tag{3.11}$$

So it's not *always* the case that knowing that a certain sentence is true goes hand in hand with knowing that sentence.

Could that be what's going on in this natural language Knower paradox? This response would have it that my epistemic situation with respect to (*) is to be described:

$$K(\ulcorner T(*)\urcorner) \wedge \neg K(*) \tag{3.12}$$

Here, though, (3.12) is not a plausible description of my epistemic situation. For while I *can* know that a certain sentence is true without knowing the sentence, I *can't* be in that position while simultaneously understanding all the terms involved in the sentence and knowing their denotations in context. (3.11) is true because I don't know what sentence John uttered after awakening. I can know that "Quarks are spin-1/2 particles" is true without knowing that quarks are spin-1/2 particles if I don't know the meanings of the constituent terms well enough to know which proposition that sentence expresses.

With (*), however, I know both the meanings and the denotations (relative to the context in which I read it) of all the terms involved. I know full well what page I'm currently reading, and I know what a starred sentence is. I know what sentence is the first starred sentence. And I know the meanings of all the terms involved—the negation sign, and the knowledge predicate. The explanation for why (3.11) could be true doesn't hold with (3.12). It's not plausible that I could know that (*) is true while failing to know it, given that I know what all the constituents of (*) mean and what sentence the definite description "the first starred sentence on the page you're reading" refers to.

Thus, (3.12) is false in this context, and we have a genuine epistemic paradox. Sentences like (*) generate genuine contradictions, assuming the factivity of knowledge, knowledge of that factivity, and single premise closure. Thus something has to give. Those who stand by classical logic must deny us some very basic knowledge—like our knowledge that we are factive with respect to (*), or knowledge that (*) is true if and only if I don't know it. Some very unpalatable bullets must be bit.

In the system I've presented, (3.12) is not simply derivable from those knowledge attributions. Embedded in the derivation of $\neg K(*)$ is a reductio step, and and on my system reductio isn't valid. You can easily deduce its negation from it, and vice versa, the hallmark of indeterminate sentences. So, on my view, your state of knowledge is indeterminate with respect to (*). This is plausible; badly behaved, semantically ungrounded sentences like (*) entail everything, and so do their negations; thus we should accept neither them nor their negations, and we certainly shouldn't reason classically with them.

The key here is that at no point is any reference made to *idealized* knowledge with mathematical content. Everything in the natural language versions of the Knower paradox can be put at the straightforward level of knowledge. Here, in order to know the all important principles of semantic ascent and descent, we don't need to imagine agents who have worked through the diagonalization theorem themselves, relative to a theory of syntax and method of diagonalization. Instead, we just need to think of agents who have looked at the page, and taken note of what starred sentences are written on it. They need perform a very small number of classical steps to arrive at contradictions. Thus the natural language knower paradox is really a paradox about knowledge proper, not about idealized knowability. You, the reader, can work competently through every literal reasoning step involved; no part of it involves a promissory ellipsis standing for a long diaganalization proof relative to a particular coding scheme, method of representing sentential predicates, and method of diagonalization. So, given the availability of directly referential terms in English like 'this', the sources of intensionality that were present in mathematical versions of the Knower paradox are avoidable.

A simplified model of how direct, natural language self-reference might work is provided in Appendix B. It's simplified in that it only accounts for rigid ways of referring to sentences, like (plausibly) 'this'. A more realistic model accommodating non-rigid terms is beyond the scope of this project, since I'm interested mainly in demonstrating the consistency of (FACT)–(KFACT)–(SPC).

3.5 Concluding thoughts: robustness, new paradoxes

There are now two broad options on the table for solving the Knower paradox. On the classical approach, the only one hitherto present in the Knower literature, we have to decide between abandoning (KFACT), and thereby implausibly limiting what agents can know, and abandoning (SPC), placing unduly strict limits on the deductive powers of agents. On my view, we can keep (FACT), (KFACT), and (SPC), at the cost of abandoning the validity of two rules of inference: unrestricted if-introduction, and unrestricted reductio.[20] How can we decide? This might be a pretty hard decision to make,

[20] And, it's worth again noting, since K_3 reduces to classical when the semantic values of sentences are 0 or 1, everyday reasoning using these rules involving exclusively sentences

if only principles concerning knowledge were at stake. But the implications are broader; principles concerning truth are also at play in the same decision. After all, anyone willing to abandon classical logic to hang on to T-out and T-in ought also be willing to use the theory for which she abandoned it to talk about knowledge, given the tight conceptual connections between knowledge and truth.

If we remain wedded to classical logic, we thereby get ourselves mired in two unsavory debates. The first unsavory debate is whether we should keep T-out, or T-in. Self-referential sentences involving a truth predicate in a classical setting seem to show that, despite all we thought we knew about how truth works, truth cannot actually work that way. Either there are instances in which p has semantic value 1 but $T(\ulcorner p \urcorner)$ doesn't, or vice-versa. Thus we need a revisionary concept of truth in order to prevent the paradox. The unsavory battle is about which way we should revise it.

By the same token, there is a different battle that we need to involve ourselves in if we hang onto classical logic in the case of knowledge. That battle is the one that has already started to take place in the literature on the Knower paradox. Are some instances of knowledge non-factive? Is it *necessarily* the case that there will be some sentences for which you cannot know the corresponding instance of the factivity schema? Are there sentences validly derivable from things that you know that you cannot know on this basis? In classical settings, we have to come down one way or another on this unsavory debate. We need to create the best second-rate knowledge predicate that we can, given the looming threat of paradoxes.

And the trouble does not stop there. Every time you have something that looks like a predicate of sentences obeying interesting structural principles, you must be on guard against paradoxes of self-reference. Indeed, nothing in the formal theory really hangs on using $\Box T(\ulcorner p \urcorner)$ to translate "p is known." We might have used the modal operator to represent alethic necessity instead. "Is necessary" seems, after all, to be a predicate. We can say things like "Some sentences are necessary, whereas others are only contingent." Thus, if we stick with classical logic, we'll have to make exactly the same choice for alethic necessity regarding (FACT), (KFACT), and (SPC) as we made for the Knower. Which is it? Some sentences are necessary but not true? It's not in general necessary that necessary sentences are true? There are counterexamples to the principle that, if you can derive a sentence from a

without semantic/epistemic vocabulary go through on my theory.

necessary sentence, then the derived sentence is necessary? I'd rather not decide.

Thus, to hang onto the classical inferences unrestrictedly, we need to have at least two distasteful debates, at the end of which we will accept in our theory at least two predicates with a less robust structure than we thought we had every reason to expect. It doesn't much matter, for these high-level purposes, which revisionary predicates ought to be accepted, if we come down on the side of classical logic; it matters for my purposes only *that* these two independent and difficult decisions must be made, if we insist on sticking with a classical logic.

The appeal of the non-classical approach is that both of these unsavory battles can be avoided. The ability to avoid both of these battles in a single go strikes me as good evidence for this kind of approach. With fairly conservative emendations to classical logic, we can keep our first-rate knowledge predicate and our first-rate truth predicate, without losing any classical reasoning in bivalent contexts. Therefore, let's do that, I say—at least until someone comes along with a defense of classical logic persuasive enough to force us to have these battles out.

Appendix A

I'll follow the basic tack taken by Caie (2012) for combining a modal language with a Kripke-style truth predicate and a Field-style conditional. We want a first-order modal language with a special predicate, T, and a special connective, $\rightarrow$. We'll think of the box operator as representing knowledge attributions—a departure from Caie, who is more concerned with paradoxes about belief. Apart from the truth-predicate and conditional, our language will have a standard-issue syntax, and for simplicity, all predicates will be one-place. We'll have no use for equality.

Start out with a countable stock of constants n_i and a countable stock of variables x_i. The syntax is generated as follows:

$$t := n_i \mid x_i$$
$$\varphi := P_i(t_j) \mid \neg\varphi \mid (\varphi\vee\varphi) \mid (\varphi\wedge\varphi) \mid \Box\varphi \mid \forall x_i\varphi \mid T(t_i) \mid (\varphi \rightarrow \varphi).$$

Models for the fragment of this language without the truth predicate or the special conditional are basically standard-issue. A base model $\mathcal{M}$ for this fragment is a quadruple $\langle D, W, R, v\rangle$. D is a non-empty set of objects, W a non-empty set of worlds, R a reflexive relation between worlds, and v an interpretation function that sends every predicate-world pair $\langle P, w\rangle$ to subsets of D and every name n_i to an element of D. We stipulate that the domain includes all well-formed formulas of the language—sentences, after all, are objects too!—and a given name is assigned to the same object in all worlds within a given model (names are rigid). A variable assignment function μ is a function from variables x_i to elements of D. Finally, there is a term-interpretation function $[\cdot]_{\mathcal{M},\mu}$, relative to a model and variable assignment (but not to a world, because names are rigid!):

$$[n_i]_{\mathcal{M},\mu} = v(n_i) \qquad \text{for constants } n_i \ ;$$
$$[x_i]_{\mathcal{M},\mu} = \mu(x_i) \qquad \text{for variables } x_i \ .$$

If you are wondering how to get self-referential sentences like the Liar and Knower out of these ingredients, see Appendix B. What's crucial for our purposes is that we can consider classes of models with some constant n_i such that $[n_i]_{\mathcal{M},\mu}$ is the sentence $\neg K(n_i)$. The Knower sentence is then the sentence $K(n_i)$. An equally paradoxical sentence in such models is the wide-scope negation version more analogous to the Liar: the wide-scope negation knower sentence is just $\neg K(n_i)$. The upshot: No classical theory can allow models with such assignments as these. The non-classical theory developed below, on the other hand, has no problem accommodating such models.

The semantics for the fragment of the language without the truth-predicate or the Field conditional is exactly what you would expect:

$[\![P(t_i)]\!]^{\mathcal{M},w,\mu} = 1$ if $[t_i]_{\mathcal{M},\mu} \in v(w, P)$, 0 otherwise;
$[\![\phi \wedge \psi]\!]^{\mathcal{M},w,\mu} = \min\{[\![\phi]\!]^{\mathcal{M},w,\mu}, [\![\psi]\!]^{\mathcal{M},w,\mu}\}$;
$[\![\phi \vee \psi]\!]^{\mathcal{M},w,\mu} = \max\{[\![\phi]\!]^{\mathcal{M},w,\mu}, [\![\psi]\!]^{\mathcal{M},w,\mu}\}$;
$[\![\neg\phi]\!]^{\mathcal{M},w,\mu} = 1 - [\![\phi]\!]^{\mathcal{M},w,\mu}$;
$[\![\Box\varphi]\!]^{\mathcal{M},w,\mu} = \min\{[\![\varphi]\!]^{\mathcal{M},v,\mu} : wRv\}$;
$[\![\forall x_i\varphi]\!]^{\mathcal{M},w,\mu} = \min\{[\![\varphi]\!]^{\mathcal{M},w,\mu'} : \mu' =_{x_i} \mu\ \}$, where $\mu' =_{x_i} \mu$ just in case μ agrees with μ' except possibly on what it assigns x_i.

A formula has a semantic value x *tout court* just in case all variable assignment functions assign it semantic value x. For sentences this condition is not hard to meet, since no change in variable assignment function can alter the semantic value. Nothing interesting happens with the quantifiers, so in what follows I'll leave out the variable assignment when talking about the semantic values of closed sentences. Similarly, when it's obvious from context that only names, and not variables, are involved, I'll leave the variable assignment function off of the term interpretation function.

The above is a typical first-order modal model. Things are different only when the special predicate T and the special connective $\to$ enter the picture. Sentences involving these elements are not guaranteed to have semantic value 0 or 1. We'll enrich a classical model $\mathcal{M}$ in two stages—the first stage gives semantic values to sentences involving T, the second stage gives semantic values to sentences involving $\to$.

First, we must enrich a given model $\mathcal{M}$ to a model $\mathcal{M}^+$ that assigns semantic values to T-involving sentences, by means of a series of models $\mathcal{M}^{+\alpha}$. All of these models will assign all formulas whose main connective is $\to$ semantic value $\frac{1}{2}$. The interpretation for T relative to a world w is a Kripke-Feferman fixed point construction. Since some sentences involving T receive semantic value $\frac{1}{2}$, we must define both an extension T^{w+} and an anti-extension for T^{w-} for T at w. We do this by means of a series of temporary extensions T^{w+}_{α} and anti-extensions T^{w-}_{α}. The final extension consists of all sentences of the language that are true (relative to the model and world); the anti-extension consists of all non-sentences, plus each sentence whose negation is true. The semantics for $\mathcal{M}^{+\alpha}$ is the same as that for $\mathcal{M}$, except for the following clause:

$$\llbracket T(t_i)\rrbracket^{\mathcal{M}^{+\alpha},w,\mu} = 1 \text{ iff } [t_i]_{\mathcal{M}^{+\alpha},\mu} \in T^{w+}_{\alpha}\ ;$$
$$\llbracket T(t_i)\rrbracket^{\mathcal{M}^{+\alpha},w,\mu} = 0 \text{ iff } [t_i]_{\mathcal{M}^{+\alpha},\mu} \in T^{w-}_{\alpha}\ ;$$
$$\llbracket T(t_i)\rrbracket^{\mathcal{M}^{+\alpha},w,\mu} = \tfrac{1}{2} \text{ otherwise.}$$

Let $T^{w+}_0 = \emptyset$, and let $T^{w+}_{\alpha+1}$ be the set of sentences φ such that $\llbracket\varphi\rrbracket^{\mathcal{M}^{+\alpha},w} = 1$. Similarly let T^{w-}_0 be the empty set, and let $T^{w-}_{\alpha+1}$ be the set of sentences φ such that $\llbracket\varphi\rrbracket^{\mathcal{M}^{+\alpha},w} = 0$, as well as all non-sentences. At a limit ordinal λ, T^{w+}_{λ} is the set of sentences φ such that, for some $\beta < \lambda$, $\llbracket T(t_i)\rrbracket^{\mathcal{M}^{+\beta},w} = 1$ (0 for T^{w-}_{λ}).

The key result for this kind of construction is the

> **Fixed Point Theorem**: For some least ordinal σ, the set of sentences φ such that $\llbracket\varphi\rrbracket^{\mathcal{M}^{+\sigma},w} = 1$ is equal to T^{w+}_{σ}.

See Field (2008) or the technical appendix of Caie (2012) for a sketch of the proof. $\mathcal{M}^+$ is simply $\mathcal{M}^{+\sigma}$ for this least ordinal σ. It remains only to say what to do with sentences involving $\rightarrow$. The idea is to build a new kind of fixed-point semantics for the conditional, using a series of things called Field models. A Field model $\mathcal{M}_\alpha$ starts with the assignments provided by $\mathcal{M}^+$. Its only real work is to assign semantic values to sentences with $\rightarrow$ as the main connective. On top of each Field model $\mathcal{M}_\alpha$ we construct a new Kripke fixed-point model $\mathcal{M}^+_\alpha$ to reclaim the right extension for T, now that some $\rightarrow$-involving sentences have just had their semantic values corrected.

The following inductive procedure determines semantic values, for all base models $\mathcal{M}$, worlds w, and variable assignment μ:

Base Field Model $\mathcal{M}_0$:

For all formulas φ,

- $[\![\varphi]\!]^{\mathcal{M}_0,w,\mu} = [\![\varphi]\!]^{\mathcal{M}^+,w,\mu}$.

Remember that this means that, for formulas of the form $\varphi \rightarrow \psi$,

- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_0,w,\mu} = \frac{1}{2}$.

For each non-limit ordinal $\alpha > 0$:

For formulas φ not of the form $\varphi \rightarrow \psi$,

- $[\![\varphi]\!]^{\mathcal{M}_\alpha,w,\mu} = [\![\varphi]\!]^{\mathcal{M}^+_{\alpha-1},w,\mu}$.

For formulas of the form $\varphi \rightarrow \psi$,

- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_\alpha,w,\mu} = 1$ iff $[\![\varphi]\!]^{\mathcal{M}^+_{\alpha-1},w,\mu} \leq [\![\psi]\!]^{\mathcal{M}^+_{\alpha-1},w,\mu}$;
- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_\alpha,w,\mu} = 0$ iff $[\![\varphi]\!]^{\mathcal{M}^+_{\alpha-1},w,\mu} > [\![\psi]\!]^{\mathcal{M}^+_{\alpha-1},w,\mu}$.

For a limit ordinal λ:

For formulas φ not of the form $\varphi \rightarrow \psi$,

- $[\![\varphi]\!]^{\mathcal{M}_\lambda,w,\mu} = x$ iff there is some ordinal $\beta < \lambda$ such that for all σ such that $\beta \leq \sigma < \lambda$, $[\![\varphi]\!]^{\mathcal{M}^+_\sigma,w,\mu} = x$;

- $[\![\varphi]\!]^{\mathcal{M}_\lambda,w,\mu} = \frac{1}{2}$ otherwise.

For formulas of the form $\varphi \rightarrow \psi$,

- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_\lambda,w,\mu} = 1$ iff there is some ordinal $\beta < \lambda$ such that for all σ such that $\beta \leq \sigma < \lambda$, $[\![\varphi]\!]^{\mathcal{M}^+_\sigma,w,\mu} \leq [\![\psi]\!]^{\mathcal{M}^+_\sigma,w,\mu}$;
- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_\lambda,w,\mu} = 0$ iff there is some ordinal $\beta < \lambda$ such that for all σ such that $\beta \leq \sigma < \lambda$, $[\![\varphi]\!]^{\mathcal{M}^+_\sigma,w,\mu} > [\![\psi]\!]^{\mathcal{M}^+_\sigma,w,\mu}$;
- $[\![\varphi \rightarrow \psi]\!]^{\mathcal{M}_\lambda,w,\mu} = \frac{1}{2}$ otherwise.[21]

Various formulas will eventually 'stabilize' at some semantic value or other (relative to a base model, world, and variable assignment function), in the sense that they retain that value for all subsequent Field models with that world and that variable assignment function. If a formula stabilizes at x, then it has Final Semantic Value x. If it does not stabilize, then it has Final Semantic Value $\frac{1}{2}$. Following Field, call this Final Semantic Value of a formula relative to a base model $\mathcal{M}$, world w in $W_\mathcal{M}$, and assignment function μ, $|||\varphi|||^{\mathcal{M},w,\mu}$. The following theorem is the final piece in the puzzle: Though various formulas may stabilize at different points in the inductive procedure, there are always future points at which *all* stabilizing formulas will have stabilized at their Final Semantic Value, *and* all non-stabilizing formulas have their rightful semantic value $\frac{1}{2}$. Field calls this the

> **Fundamental Theorem**: For all ordinals σ there's some ordinal $\eta > \sigma$ such that, for every formula φ, variable assignment function μ, and world w, if $|||\varphi|||^{\mathcal{M},w,\mu} = x$ then $[\![\varphi]\!]^{\mathcal{M}^+_\eta,w,\mu} = x$.

Field calls these ordinals "acceptable." Validity within the class of acceptable Field models (i.e. a Field model M_α for an acceptable ordinal α) is pretty much what you would expect. For the class $\mathbb{M}$ of all acceptable Field models and sentences φ and ψ, $\varphi \models_\mathbb{M} \psi$ just in case, for every $\mathcal{M} \in \mathbb{M}$ and $w \in W_\mathcal{M}$, if $|||\varphi|||^{\mathcal{M},w} = 1$ then $|||\psi|||^{\mathcal{M},w} = 1$. With this definition of

[21] The limit ordinal definition trivially covers the non-limit-ordinal definition too, but it's more perspicuous to think about the two cases separately, because the two cases result in different kinds of behavior. Roughly speaking, for badly-behaved sentences involving $\rightarrow$ (like the Curry sentence, discussed below), the non-limit ordinals result in the semantic value of the sentence oscillating between 1 and 0 at successive steps, and at a limit ordinal, they find a temporary respite from this oscillation at $\frac{1}{2}$.

validity, the Fundamental Theorem guarantees that the corresponding logic remains K_3 and not something weird: all individual Kripke-plus-Field models have a K_3 logic, and the formulas are all together stabilized at the acceptable Kripke-plus-Field models.

The full logic for the conditional (excluding interesting mixes of modal principles with the conditional) is found in Field (2003), p. 292. The important results for our purposes are simply these: In bivalent contexts (i.e. those for which excluded middle $\varphi \vee \neg\varphi$ holds), $\rightarrow$ and $\supset$ are equivalent. Modus ponens is valid for $\rightarrow$ unrestrictedly.

The main classical rule that is *not* valid is if-introduction: It may be that $\varphi \models_{\mathbb{M}} \psi$, but $\varphi \rightarrow \psi$ does not have Final Semantic Value 1 for some worlds and models. This is actually good news: the addition of classical if-introduction into a logic that has the unrestricted T-schema and modus ponens renders the theory inconsistent, because of Curry's paradox, a Liar-like paradox of self-reference that centers on the sentence $c := T(\ulcorner c \urcorner) \rightarrow \bot$.

On this construction, $T(\ulcorner c \urcorner) \rightarrow \bot \models_{\mathbb{M}} \bot$. Why? $T(\ulcorner c \urcorner) \rightarrow \bot \models_{\mathbb{M}} T(\ulcorner c \urcorner)$ because T-in is valid, and $T(\ulcorner c \urcorner), T(\ulcorner c \urcorner) \rightarrow \bot \models_{\mathbb{M}} \bot$ because modus ponens is valid. Nonetheless, $(T(\ulcorner c \urcorner) \rightarrow \bot) \rightarrow \bot$ has Final Semantic Value $\frac{1}{2}$: the consequent stabilizes at semantic value 0, but the antecedent does not stabilize. The antecedent is just the Curry sentence. At some stages in the Field construction, this will have semantic value 1; at the next stage the main $\rightarrow$ therefore gets semantic value 0. But at subsequent stages the conditional in the Curry sentence gets semantic value 0, which gives the main $\rightarrow$ semantic value 1. This oscillation continues *ad infinitum*, so the whole sentence receives Final Semantic Value $\frac{1}{2}$.[22]

Finally, here are the proofs of propositions 4.1, 4.2, and 4.3:

Proposition 4.1. *All instances of the following schema are valid:*

$$\Box T(\ulcorner \varphi \urcorner) \rightarrow \varphi.$$

Proof. At a world w, $\Box T(\ulcorner \varphi \urcorner)$ has the minimum semantic value of $T(\ulcorner \varphi \urcorner)$ at all worlds accessible from w, and the relation R is reflexive. Therefore the minimum semantic value for $T(\ulcorner \varphi \urcorner)$ at all worlds accessible from w

[22] See Field (2008), §4.1 for a more involved discussion of Curry's paradox. I'll just note that this is enough to guarantee that the Field conditional is not truth functional: Some $\frac{1}{2} \rightarrow \frac{1}{2}$ sentences, for example $T(\ulcorner l \urcorner) \rightarrow T(\ulcorner l \urcorner)$, have semantic value 1, whereas others, like $T(\ulcorner l \urcorner) \rightarrow T(\ulcorner c \urcorner)$, have semantic value $\frac{1}{2}$.

cannot be strictly greater than that of $T(\ulcorner\varphi\urcorner)$ in w. But at any world in an acceptable Kripke-plus-Field model, $T(\ulcorner\varphi\urcorner)$ has the semantic value of φ. Therefore the semantic value of $\Box T(\ulcorner\varphi\urcorner)$ is less than or equal to that of φ for all Kripke-plus-Field models. Therefore the semantic value of $\Box T(\ulcorner\varphi\urcorner) \to \varphi$ stabilizes at 1 for all worlds and models. Thus $\models_{\mathbb{M}} K(\ulcorner\varphi\urcorner) \to \varphi$. □

Proposition 4.2. *All instances of the following schema are valid:*

$$\Box T(\ulcorner \Box T(\ulcorner\varphi\urcorner) \to \varphi\urcorner).$$

Proof. As we saw in the text, Proposition 4.1 and Proposition 4.3 together entail each instance of this axiom. □

Proposition 4.3. *The following rule is valid:*

$$\begin{array}{|l l}
\Box T(\ulcorner\varphi\urcorner) & \\
\begin{array}{|l} \varphi \\ \hline \dots \\ \psi \end{array} & \text{(Only axioms and rules of K}_3\text{FT.)} \\
\Box T(\ulcorner\psi\urcorner) & \text{Closure}_{K_3FT}.
\end{array}$$

Proof. Suppose that $\Box T(\ulcorner\varphi\urcorner)$ has semantic value 1 at some arbitrary world w in some arbitrary Field model $\mathcal{M}$, and suppose $\varphi \vdash_{K_3FT} \psi$. Since (exercise to the reader!) the rest of K_3FT is sound on $\mathbb{M}$, it follows that $\varphi \models_{\mathbb{M}} \psi$. We just need to show that $\Box T(\ulcorner\psi\urcorner)$ has semantic value 1 at w in $\mathcal{M}$. $\varphi \models_{\mathbb{M}} \psi$ holds just in case, for all models and worlds, if φ has Final Semantic Value 1, so does ψ. $\Box T(\ulcorner\varphi\urcorner)$ has value 1 just in case all accessible worlds from w give $T(\ulcorner\varphi\urcorner)$ semantic value 1. Since **T-out** and **T-in** are both valid on Field's construction, if $T(\ulcorner\varphi\urcorner)$ has semantic value 1 at w in $\mathcal{M}$, so does φ. But since $\varphi \models_{\mathbb{M}} \psi$, all worlds that give φ semantic value 1 also give ψ semantic value 1, and all worlds that give ψ semantic value 1 also give $T(\ulcorner\psi\urcorner)$ semantic value 1. Therefore, in all worlds accessible from w, $T(\ulcorner\psi\urcorner)$ has semantic value 1. So the minimum semantic value of $T(\ulcorner\psi\urcorner)$ at all worlds accessible from w is 1; but that means that $\Box T(\ulcorner\psi\urcorner)$ has semantic value 1 at w. This holds for all worlds and models, so this rule is valid: For all models and worlds, if $\Box T(\ulcorner\varphi\urcorner)$ has semantic value 1 and $\varphi \vdash_{K_3FT} \psi$, then $\Box T(\ulcorner\psi\urcorner)$ also has semantic value 1.[23] □

APPENDIX B

You may have been wondering about how exactly to get self-referential sentences like the Liar and Knower out of these ingredients. Self-reference can be achieved in one of two ways: the cheap way, and the honest way. The honest way is to put a bit of arithmetic into this theory. Field makes an "important observation" in §1.1 of Field (2008), according to which the results of the diagonalization theorem hold even in theories whose logic is weaker than classical, provided only that classical logic holds for the arithmetic portion of the language, standard quantifier reasoning is allowed, and the logic of the bi-conditional is minimally reasonable.[24] It would be a routine matter to plop enough arithmetic into the theory above to profit from the results of the diagonalization lemma, even with a weaker logic. But I won't be honest for these purposes.

Instead, I'll achieve self-reference in the cheap way. The cheap way is to focus directly on which terms denote which sentences in the models. The Liar sentence, for example, would exist in any model where a name n_i denoted the sentence $\neg T(n_i)$. A name n_i is a **Liar name** in model $\mathcal{M}$ just in case $[n_i]_{\mathcal{M}} = \neg T(n_i)$. $\neg T(n_i)$ is the Liar sentence.

We interpret the corner quotes differently when we're smuggling self-reference in on the cheap. When we earned our self-reference via the diagonalization lemma in PA, $\ulcorner\varphi\urcorner$ was a metalinguistic name that stood for the Gödel number of the formula φ. Here, $\ulcorner\varphi\urcorner$ is still a metalinguistic name, but it stands, not for a Gödel number, but rather for the name n_i that denotes the formula φ. Since we want $\ulcorner\cdot\urcorner$ always to be well-defined, we saddle our models with the following restriction: Fix a particular enumeration of the

[23] Also $K(\ulcorner\varphi \to \psi\urcorner), K(\ulcorner\varphi\urcorner) \models_{\mathbb{M}} K(\ulcorner\psi\urcorner)$ for similar reasons. However, the inference from $\varphi \vdash \psi$ to $K(\ulcorner\varphi\urcorner) \to K(\ulcorner\psi\urcorner)$ is not valid—but only for Curry-paradox-related reasons concerning conditional introduction. Generally, multi-premise logical closure is valid on this construction: $\varphi_1, \ldots, \varphi_n \vdash_{K_3FT} \psi$ entails $K(\ulcorner\varphi_1\urcorner), \ldots, K(\ulcorner\varphi_n\urcorner) \vdash_{K_3FT} K(\ulcorner\psi\urcorner)$.

[24] "Minimally reasonable" means: $A \leftrightarrow A$ and $\exists x[x = t \wedge C(x)] \leftrightarrow C(t)$ are theorems, and if $A \leftrightarrow B$ is a theorem then substituting A for B preserves theorem-hood. The Field conditional validates both of these requirements. Of course, even to state these conditions, we'd need to add equality to our language, which hitherto I, following Wittgenstein, have eschewed.

names. We insist that, in every model, every sentence has exactly one name, and that each sentence gets the same name in every world of every model.[25] Since there are countably many names, and countably many sentences, we have enough names to go around.[26]

Classical theories with a sentential truth predicate that satisfies the semantic equivalence of φ and $T(\ulcorner\varphi\urcorner)$ *cannot* admit Liar names into models, on pain of inconsistency. That is, if the semantics winds up obeying

$$[\![T(\ulcorner\varphi\urcorner)]\!]^{\mathcal{M},w} = [\![\varphi]\!]^{\mathcal{M},w}, \qquad \text{T-equiv}$$

then there can be no model $\mathcal{M}$ in which $[n_i]_{\mathcal{M}} = \neg T(n_i)$. If it did, then

$$\begin{aligned}
[\![\neg T(n_i)]\!]^{\mathcal{M},w} &= 1 - [\![T(n_i)]\!]^{\mathcal{M},w} && \text{Semantics of } \neg;\\
&= 1 - [\![[n_i]_{\mathcal{M}}]\!]^{\mathcal{M},w} && \text{By T-equiv;}\\
&= 1 - [\![\neg T(n_i)]\!]^{\mathcal{M},w} && \text{By what } n_i \text{ denotes.}
\end{aligned}$$

Thus the Liar cannot consistently be assigned a semantic value: if it's 0, it's 1, and vice-versa. So even classical theories *without* the arithmetic needed for the diagonalization lemma must nonetheless place ad-hoc restrictions on which names can denote which objects. With enough arithmetic in hand, self-reference (though of a slightly different sort, relying on the provability of certain biconditionals rather than focusing directly on which terms denote which sentences) becomes unavoidable even by stipulation.

The Knower sentence is a slightly fancier Liar-like sentence. A name n_i is a **Knower name** in model $\mathcal{M}$ just in case $[n_i]_{\mathcal{M},w} = \neg\Box T(n_i)$. The Knower

[25] These assumptions are wildly unrealistic, and mostly introduced for simplicity.

[26] This difference in our interpretation of the corner quotes will change what the actual proofs in the object language look like, once the corner quotes are interpreted. When we're doing Gödel numbering, the crucial lines in the proof where self-reference gets its bearings is in the provable biconditionals: $l \equiv \neg T(\ulcorner l \urcorner)$, where the l in question is actually some complicated arithmetic formula, and $\ulcorner l \urcorner$ is some natural number. When we're smuggling in self-reference on the cheap, these biconditionals will have the form of mere propositional tautologies: l is really just $\neg T(n_i)$ where n_i is a Liar name, and since $\ulcorner\cdot\urcorner$ just takes a formula to its name, the right hand side of the equivalence is also $\neg T(n_i)$. The force of self-reference comes not from these biconditionals, which are tautologies, but rather from the T-schema. The instantiation of T-out for the Liar sentence, for example, will have the form: $T(n_i) \supset \neg T(n_i)$. Thus, with either way of achieving self-reference, the proofs written with the metalinguistic corner quotes look exactly the same; but when you interpret the corner quotes and look at the honest-to-God object-language forms of the proof, they look a bit different.

sentence (with narrow-scope negation) is then $\Box T(n_i)$. This sentence "says" that its own negation is known. The same kind of reasoning shows that classical models satisfying the semantic equivalence of $T(\ulcorner\varphi\urcorner)$ with φ cannot admit Knower names:

$$\begin{aligned}
[\![\Box T(n_i)]\!]^{\mathcal{M},w} &= Min\{[\![T(n_i)]\!]^{\mathcal{M},w'} : wRw'\} && \\
&\leq [\![T(n_i)]\!]^{\mathcal{M},w} && \text{Because } wRw;\\
&= [\![[n_i]_{\mathcal{M}}]\!]^{\mathcal{M},w} && \text{By T-equiv;}\\
&= [\![\neg\Box T(n_i)]\!]^{\mathcal{M},w} && \text{By what } n_i \text{ denotes;}\\
&= 1 - [\![\Box T(n_i)]\!]^{\mathcal{M},w} && \text{Semantics of } \neg.
\end{aligned}$$

Thus $[\![\Box T(n_i)]\!]^{\mathcal{M},w}$ must be 0, since if it's 1 then $1 \leq 1 - 1$. So $[\![\neg\Box T(n_i)]\!]^{\mathcal{M},w} = 1$. That means that there is some w' accessible from w at which $[\![\neg T(n_i)]\!]^{\mathcal{M},w'} = 1$. Now, since w was arbitrary in the above reasoning, the same argument as above shows that $[\![\neg\Box T(n_i)]\!]^{\mathcal{M},w'} = 1$. But:

$$\begin{aligned}
[\![\neg\Box T(n_i)]\!]^{\mathcal{M},w'} &= [\![[n_i]_{\mathcal{M}}]\!]^{\mathcal{M},w'} && \text{Because of what } n_i \text{ denotes;}\\
&= [\![T(n_i)]\!]^{\mathcal{M},w'} && \text{By T-equiv.}
\end{aligned}$$

So at w', $T(n_i)$ must be assigned semantic value 1. But w' was introduced precisely to witness w's accessibility to a $\neg T(n_i)$ world, which would require $T(n_i)$ to be assigned semantic value 0 at w'. Thus no classical models satisfying T-equiv can include Knower names.

This argument also shows that the wide-scope negation Knower sentence used by Maitzen (1998) cannot appear in classical models. That sentence uses the same n_i as above, but is the very sentence that n_i denotes: $\neg\Box T(n_i)$. The above argument shows that Knower-names are forbidden from classical models; therefore, the wide-scope and narrow-scope negation Knower sentences are equally forbidden.

Chapter 4

Paradoxical Desires

4.1 The paradox

There's almost nothing you can't want. A glass of Bordeaux, a degree in astrophysics, world peace ... You name it, and there's probably someone, somewhere, who wants it.

It often happens that our desires involve other people's desires. Jennifer Aniston doesn't just want her deadbeat boyfriend in the 2006 classic *The Break Up* to do the dishes; she wants him to *want* to do the dishes. Conversely, our desires can involve a blind kind of reference to others' desires. I may not particularly care where we go to dinner, but care very much about the enjoyment of my more opinionated dining comrades. Thus I might want simply to go wherever my comrades most want to go. Similarly a parent might want to get their daughter whatever she most wants for her birthday, without having any idea what that is. Desire-directed desires such as these abound.

That isn't to say that they can't get us into trouble. Imagine that my dining comrades turn out to be as unopinionated about where to dine as I am. Then we might end up in the following sort of situation:

> **My strongest desire**: That we dine wherever my dining comrades most want to dine.
>
> **My comrades' strongest desire**: That we dine wherever *I* most want to dine.

In this case, we've got a problem on our hands: What I want depends on what my comrades want, and vice-versa, and we're stuck. It will be completely unclear where we should dine until some of us change our desires. (My colleagues and I often languish indecisively in such states.)

We are well-advised to avoid desires such as these when they aren't supplemented by other, more basic desires. They are the cause of many indecisive wallowings among domestic partners, friends, and colleagues. But the situation can get far worse than indecisive strife; desire-directed desires can lead to outright paradox.

Here, it seems, is a perfectly possible pair of desires that could be had by two acquaintances, Mal (think 'malevolent') and Ben (think 'benevolent'):

> **Mal's strongest desire**: That Ben doesn't get whatever he most strongly desires.
>
> **Ben's strongest desire**: That Mal gets whatever she most strongly desires.

Cases like this surely occur less frequently than the cooperative predicament. We require Mal to harbor a particular kind of ill-will towards Ben, who in turn has nothing but benevolent desires towards Mal. We cannot expect Ben to endure such treatment. But endure it he might. Perhaps Ben is a religious man, albeit a hedonistic one, whose life purpose is to bring about the satisfaction of others' desires. Perhaps Mal is a misanthrope, who wants all benevolent men such as Ben to have their sanctimoniously benevolent desires frustrated. Whatever we say about this case, it represents a *prima facie* possible pair of desires. Mal and Ben might have existed.

Those familiar with the Liar paradox (especially the contingent versions presented in Kripke (1975)) will have presaged a paradox. And indeed, one looms. The problem in the case of cooperative desires was that nothing would happen. In order to know what I want, we have to figure out what my dining comrades want; but in order to know what they want, we have to figure out what I want. And so nothing gets done. A regrettable outcome, but a perfectly coherent one.

In the case of desires like Mal's and Ben's, however, we risk encountering far bigger game: true contradictions. When you desire something, it seems, one of two things can happen. You either eventually get what you want, or you fail to. So we can ask in this situation: whose strongest desire is satisfied here? Mal's? Ben's? Both? Neither? A bit of reasoning shows that none of these answers makes any sense.

Here's the proof. Suppose that Mal gets what she most strongly desires. That just means that Ben doesn't get what he most strongly desires. But Ben simply desired that Mal's strongest desire be satisfied. Since Ben doesn't get what he most desires, neither can Mal. But we supposed that Mal *does* eventually get what she most strongly desires. Thus this assumption must be false; it must be that Mal *doesn't* eventually have her strongest desire satisfied.

However, this hypothesis is no better! Mal most strongly desires Ben's strongest desire to be frustrated; so if she doesn't get what she desires, that can only be because Ben gets what he desires. But since Ben just wants Mal to get what *she* desires, the satisfaction of Ben's desire requires the satisfaction of Mal's. Thus, supposing that Mal's strongest desire is frustrated, we can prove that it's satisfied. Contradiction.

Here's another way to see the contradiction: Because of what Mal wants, Mal's desire is satisfied if and only if Ben's isn't ($m \leftrightarrow \neg b$). But because of what Ben wants, Ben's desire is satisfied if and only if Mal's is ($b \leftrightarrow m$). Therefore Ben's desire is satisfied if and only if it isn't satisfied ($b \leftrightarrow \neg b$), a classical contradiction.

It's important to point out that this is a putative contradiction in our description of the world, not merely in the contents of Mal's and Ben's desires. It's a common idea that I can have incompatible desires: I might want to drink a third glass of wine (because it will taste pleasant), and simultaneously want not to drink it (so as to avoid subsequent hangovers).[1] But this case is different: the contradiction concerns whose desire is, in fact, satisfied. Mal's and Ben's desires are each perfectly internally coherent, in the sense that there is a possible state of the world in which each (individually) is satisfied. Mal's strongest desire would be satisfied, for example, in a world in which Ben most strongly desired a beer and failed to get a beer. But in *our* world, in which Mal and Ben have the desires ascribed above, a plain contradiction follows from any classical way of saying whose strongest desire is satisfied.

4.2 Responses

What does this paradox show? Two responses naturally suggest themselves:

[1] See Phillips-Brown (2017) for more on these internally inconsistent desire attributions.

> Response 1: Deny the possibility of the case. We could hold that the case presents merely *prima facie* possible pairs of desires; actually this description misrepresents the underlying psychological reality.
>
> Response 2: Accept the possibility of the case, but try to block the reasoning that leads to contradiction.

Response 2 comes in two kinds:

> Response 2a: Deny a principle of desire satisfaction, particularly the principle that, if *S* most strongly desires that *p*, then that *S*'s strongest desire is satisfied if and only if *p*.
>
> Response 2b: Deny a logical principle—most naturally, either bivalence or non-contradiction.

In what follows, I'll argue that Response 1 and 2a aren't right, and so some version of Response 2b must be. I conclude by drawing out some consequences. In particular, cases like this undermine the assumption, prevalent in the literature on the Liar paradox, that paradoxes of self-reference (broadly construed) pose a problem peculiar to language, and in particular for the expressive power of the truth predicate. If cases like the one I've described are possible, then structurally identical paradoxes arise at the level of thought itself, regardless of the expressive resources of the language in which we express those thoughts. This, I'll argue, provides a new reason for adopting a broadly non-classical approach to semantic paradoxes more generally, because solutions that hold onto classical logic by denying the T-schema are not applicable to these non-linguistic versions. But first: can cases like this even arise?

4.3 Denying the case

The above proof purports to show that the arrangement of the world putatively described above, though seemingly possible, cannot obtain. Mal and Ben can't mutually desire what they seem to desire.

It's one thing simply to stipulate this; it's another to explain why it is so. What, beyond a non-explanatory inconsistency proof, explains *why* Mal and Ben can't have the desires that they seem for all the world to have?

A proponent of Response 1 has a choice to make here. She can deny the possibility of desires like Mal's and Ben's intrinsically, or relationally. The intrinsic version of Response 1 denies that anyone can *ever* have desires of the forms that Mal's and Ben's seem to have, regardless of who else desires what. The relational version allows that we can sometimes have desires like those, but holds that when certain global features of the world obtain (features concerning who else desires what), we can no longer have those desires. I'll argue that each of these broad kinds of approach faces insuperable problems, and that neither satisfactorily explains why the Mal/Ben case can't obtain.

4.4 Intrinsic case-denying

According to the intrinsic version of Response 1, we can just never have desires like Mal's and Ben's, regardless of which desires other people happen to have. Indeed, to fully block the possibility of paradoxes like the one above, we'd have to ensure that we never really have desires whose content inextricably involves the satisfaction or non-satisfaction of others' desires. Other, more complicated paradoxes loom, and pretty much any local, desire-satisfaction-involving kind of desire can be turned into a paradoxical one with enough case-rejiggering.

The challenge for this way of responding is to offer some story about why we can't have desires like these, even in innocuous cases. As motivated above, desire-involving desires are commonly attributed in everyday life. So someone who wanted to deny these desires intrinsically would have to give a theory of content that provided some satisfactory error theory for these attributions.

What are the prospects for such a theory? Not good, I think. In recent literature on propositional attitudes and mental content, there are two broad ways that people have theorised about beliefs and desires, and what sorts of contents those states have. On some views, like those of Geach (1957), Sellars and Chisholm (1957), Dummett (1974), Davidson (1975), Fodor (1975), and Field (1978), propositional attitudes are essentially language-like in nature. Believing a proposition is standing in a relation to a sentence in what's sometimes called a 'language of thought', and this language has many of the features (compositionality, for example) that we associate with ordinary natural languages.

According to an alternate picture, most prominently advocated by Lewis

(1979) and Stalnaker (1984), beliefs and desires aren't linguistic in structure. Instead, these states are conceived of pragmatically, as fundamentally tied less to the assertion of sentences than to the explanation and prediction of rational action. On many views of this kind, propositions are modeled as something less fine-grained than sentences—sets of possible worlds, or centered worlds, or something like that—together with some sort of story about how relations of believing and desiring to these sorts of things can play a role in the explanatory, folk-psychological theory that is their home.

On neither of these ways of thinking about content is there any good reason to think that desires like Mal's and Ben's can't arise intrinsically.

On the language of thought picture, to have a belief or desire is to stand in a certain kind of relation to some sentence-like object. Which sentences in the language of thought can I believe or desire? Any one, proponents of this view generally think, that can be understood or asserted by the person doing the believing or desiring. Now, we can definitely have *some* thoughts about other peoples' desires. There's no motivation for thinking that we can't *believe* that someone's given desire is satisfied or not. We definitely have the concept of desire. But given that we posses this concept, denying the existence of desire-involving desires seems just as unmotivated as denying the existence of truth-predicate-involving sentences. Since we can entertain some sentences in the language of thought which involve the notion of desire, what prevents us from ever standing in the relation of desiring to them?

Denying that we can desire things about desires would be an ad-hoc and unmotivated constraint from the language of thought perspective. *Why*, beyond the fact that it can lead to Mal/Ben-like situations, did we evolve such that certain language-of-thought sentences can go in the belief-box but never in the desire-box? If anything, desire seems less constrained than belief, not more. The language-of-thought proponent who wanted to deny the case intrinsically would need to cook up some sort of story about why we are forbidden from desiring certain language-of-thought sentences that we can certainly believe, and I can't see a promising way to work it out.

Indeed, Whittle (2017) has shown an impossibility result in this vicinity. He's shown that *if* propositions are structured, as the language-of-thought theorist insists, then self-referential (and hence paradoxical) propositions are unavoidable. He shows how a propositional version of the diagonalization lemma can be proved with very minimal assumptions about how propositions would have to be structured, if they are structured at all.[2] So the only real hope for denying the existence of paradoxical contents (of which the Mal/Ben case is a contingent version) lies with an unstructured theory of content.

On broadly Stalnakerian views, there's no commitment to the idea that propositions have a sentence-like structure that corresponds in any straightforward way to that of the sentences we assert. Instead, thoughts are conceived of as those things the attribution of which can explain why rational agents act in certain ways. So we might attribute to someone a desire to stay dry and a belief that it's raining as part of an explanation as to why they grabbed an umbrella on the way out the door. It's this role that desire/belief attributions are supposed to serve, and they are given only as much structure as required to serve that role. Since it's hard to see how, for example, a belief in $p \wedge q$ could contribute anything different to a rational agent's actions than a belief in $q \wedge p$, many have thought that this structure will be something less fine-grained than that of sentences.

This picture is more promising than the language-of-thought one for denying the existence of desire-involving desires, for this view connects the states of believing and desiring less tightly to language. While we can pretty clearly form *sentences* about someone's desiring something, and about that desire's being satisfied or unsatisfied, the Stalnakerian doesn't assume that this maps in any isomorphic way to the structure of thought. Instead, thought has only the amount of structure required to explain rational action.

Can non-structured theories make a plausible case for denying desire-involving desires? I don't think they can. Even on Stalnaker's own terms, we cannot get away with forbidding these kinds of desires. The reason is that complicated enough profiles of rational, desire-directed actions can *necessitate* these kinds of attributions. Mal might be acting in ways that make the attribution of Ben-desire-involving desires explanatorily indispensable. Imagine that you observe Mal exhibiting the following behavior. Whenever she sees Ben reaching for something, she swats it away. Whenever Ben applies for a job, Mal destroys the application. Whenever Ben says that he wants something, Mal does everything in her power to prevent its being brought about. In all the nearby counterfactual situations in which Ben has certain desires, Mal is lurking, trying to frustrate them.

[2] The diagonalization lemma is the tool used to generate analogues of self-referential sentences in sufficiently rich mathematical languages. See Boolos, Burgess, and Jeffrey (2007) for details.

If that's how Mal is behaving, there doesn't seem to be any way to describe her state of mind other than with a Ben-desire-involving desire. Say that, in the actual world, Ben happens to desire ice cream. In this world, it's true that Mal desires that Ben not have ice cream. But *just* to say this misses out on an important counterfactually robust feature of Mal's mental state. In a nearby situation in which Ben starts wanting macaroons instead, the theorist who attributed to Mal a desire that Ben not have ice cream won't make as good of predictions as those who attributed to Mal a desire that Ben not get what he most strongly desires. So, even on a Stalnakerian picture, we shouldn't think that desire-satisfaction-involving desires never occur. The only hope for denying the Mal/Ben case is to say that something goes wrong, not with Mal's desire on its own or Ben's desire on its own, but rather with the paradoxical relation in which they happen to stand.

4.5 Relational case-denying

What about the relational version of Response 1? According to this kind of view, it's possible for Ben to have his desire, and it's possible for Mal to have her desire. But what's *not* possible is for Mal and Ben to have these desires together. Let's fix Mal's desire, and say that, initially, their desires are these:

> **Mal's strongest desire**: That Ben doesn't get what Ben most strongly desires.
>
> **Ben's strongest desire**: That Monika get the job she applied for.

The relational proponent of Response 1 need not deny that this is a perfectly possible situation. The result is that Mal winds up with a desire that (given what Ben desires) is satisfied just in case Monika does not get the job she applied for.

Now suppose that, in this situation, Ben forgets all about Monika, and reflects upon his positive feelings towards Mal. 'What a fine person Mal is!' Ben thinks; 'I really just hope that she gets whatever it is she most wants.'

The question for the relational case-denier is, what should we say about what happens to Mal? The answer, in any case, is going to be strange. Either Ben's forming this desire immediately robs Mal of her previous desire, without changing anything about her, or it prevents Ben from getting into the state of desire he'd otherwise have gotten into if Mal's strongest desire had simply been to drink a nice glass of wine. (Or both.)

Is any of these options plausible? Any of these options involves a commitment to a pretty radical kind of content externalism—too radical, I think, to stomach. Since Putnam (1981) it's been a commonly held idea that which propositions I believe and desire supervenes on more than my own internal psychological state; it depends partly on features of the world. So here on Earth, I have beliefs and desires involving H2O, but on Twin Earth that exact same internal psychological state would put me into belief and desire relations to XYZ instead. In a similar vein, Burge (1979a) argues that the content of my beliefs about arthritis is fixed by more than just my own private understanding of how arthritis works; instead, some such facts (like whether it can occur in my thigh) are fixed by relevant experts. So my belief that I have arthritis in my thigh is false, even though it might be true according to my own internal conceptual scheme.

According to the relational case-denier, something similar is happening here. A single intrinsic psychological state can, depending on the rest of the world, put Mal into different desiderative states. If Ben has normal desires, it puts Mal into a Ben-desire-involving desiderative state; if Ben is in the state described in the original case, this psychological state puts Mal into…well, some other kind of state. (It would be a burden on this kind of theorist to say exactly what Mal's desiderative state is in this case.)

Though externalism about mental content is popular these days—indeed, I count myself as an externalist, largely for Putnam-Burge style reasons—I don't think we should be happy with *this* kind of externalism. Why not? Well, the traditional defenders of content externalism do not simply assert that content supervenes widely, and leave things at that. All such cases come together with a very natural explanation for *why* and *how* our beliefs/desires can differ in content without our differing in intrinsic psychological state. In the case of Twin Earth, Putnam doesn't simply assert that the content of water-like beliefs differs; he gives an account that explains why this is so, and why it makes sense to use a notion of content that behaves that way. The reason why I can't have thoughts about XYZ, but can have thoughts about H2O, is that I've never been in the right kind of causal contact with XYZ to have thoughts about it. It's this causal story of how representational states get their content that explains the wide supervenience in Twin Earth cases.

Similarly for Burge. We have a notion of content that allows for deference to experts because we live in a social community that benefits from a division of intellectual labor regarding the understanding of various phenomena. Doctors get paid to know about arthritis and where it can occur, and I defer to them conceptually because they know about the phenomenon in question better than I do. Their expertise, and my deference to it, naturally explains why and how content supervenes widely in the way that it does.

It's really hard to see what a similar kind of explanation would look like in the Mal/Ben case. Mal and Ben appear to have all the concepts necessary to entertain the thoughts they seem to be entertaining. They are perfectly acquainted with each other, and both of them have the concept of desire. Both of them know what it is for a desire to be satisfied or not. So they have all of the concepts and acquaintances with the right objects in order to have the thoughts/desires they seem to have. The *only* reason to think that they can't have those desires in this global case is that they are (classically) mutually inconsistent; we have as yet no theory that explains why the content of our desires can depend on the whims of other people in this particular way. And, unlike in the case of H2O and XYZ, we have no good story about how even to describe Mal's and Ben's desiderative states instead.

So it's better to allow that Ben and Mal can have these desires, mutually, in this case. That only leaves the pesky little problem of the paradox. What's going on, if we seem to have a proof that this situation, which we have no good independent reason to think can't arise, can't arise?

4.6 Denying a principle of desire-satisfaction

The only hope for retaining classical logic while holding onto the possibility of the Mal/Ben case lies in denying a non-logical principle that goes into the derivation of a contradiction. Let us therefore look more closely at how this contradiction was derived.

One way was via the biconditionals $m \leftrightarrow \neg b$ and $b \leftrightarrow m$. These biconditionals each have intuitive plausibility, even in the paradoxical case. But this intuitive plausibility can be bolstered; they are derivable from seemingly incontrovertible principles relating desires to their conditions of satisfaction. The thought is that Mal's state of desire gives rise to, and indeed is characterised by, the first biconditional. What Mal wants is for Ben's desire to be frustrated; to give her what she wants is just to frustrate Ben's desire, and to fail to give her what she wants is just to satisfy Ben's desire. So her desire is satisfied just in case Ben's isn't. Lying behind this thought is a principle of desire satisfaction:

> **Desire satisfaction schema (DSS)**: If S desires that p, then S's desire that p is satisfied if and only if p.

The relevant instance is:

> If Mal desires that Ben's strongest desire be frustrated, then her desire that Ben's strongest desire be frustrated is satisfied if and only if Ben's strongest desire is frustrated.

This principle suffices to generate, in classical logic applied to the Mal/Ben case, the biconditional $m \leftrightarrow \neg b$. For Mal's strongest desire *is* that Ben's strongest desire fail to obtain. Thus, by (DSS), it is satisfied if and only if Ben's strongest desire is frustrated.

(DSS) has intuitive plausibility. But there is good reason to doubt that it is always true. Desire reports are generally thought to contain normality presuppositions as part of their meaning. So, imagine that John wants to drink a beer, and you give him a poisoned beer. He drinks it and dies. Here, it seems like it's true that John drinks a beer; and it's true that John desired to drink a beer; but plausibly, John does not get what he wants.

You might wonder if this has something to do with the indefinite article: John's real desire is to get a [normal, non-poisoned] beer. But this behavior seems to persist even without indefinite articles. Suppose John forms a desire to drink *this* beer in front of him, not knowing that his enemies have poisoned it. (DSS) commits us to:

> John's desire to drink this beer is satisfied iff John drinks this beer.

Here, intuitions may not be as clear. It seems like there is *some* sense in which, if John drinks this beer, his desire is satisfied. It's just not satisfied in the way he wanted it to be satisfied—the normal, refreshing, non-poisoned way. This way of describing things seems relatively natural, and it does not involve denying (DSS). John's desire to drink this beer is satisfied, just not how he would have liked.

But there is another temptation to say that John doesn't actually get what he wants. After all, drinking the poisoned beer is really bad for John; if John knew that the beer were poisoned, he would immediately renounce his desire for it, and would have *no* inclination to drink it. So it's odd to say that he gets what he wants in spite of his untimely and undesired demise.[3]

I feel the pull of both intuitions. Thus, I won't rely on (DSS) in my argument against this way of blocking the paradox. For the desire satisfaction principle *restricted to strongest desires* is much more plausible. That principle says:

> **Strongest desire satisfaction schema (SDSS)**: If *S* most strongly desires that *p*, then *S*'s strongest desire that *p* is satisfied if and only if *p*.

This schema is not subject to the objections to the fully general (DSS). If John claims most strongly to desire a beer, receives a poisoned one, and protests that he didn't get what he wanted, it would be reasonable to retort that he must not have *most strongly* desired a beer after all.[4] His subsequent protestation shows that he had another, logically stronger desire all along: a desire for a non-poisoned beer. Desires may be the sorts of things that can fail to be satisfied, even while their content comes true. It may, after all, come true in deviant ways—ways inconsistent with stronger desires that the agent has. But strongest desires aren't like that. Dissatisfaction with deviant ways of satisfying them merely show that they weren't actually strongest desires after all.

Perhaps there are lingering doubts about this. Perhaps even logically strongest desires have deviant ways of failing to be satisfied, such that it could make sense to say that someone most strongly desired *p*, and *p* was indeed the case, but they failed to get what they most strongly desired. Even if so, related paradoxes loom for which such a response wouldn't apply. For

[3] The kind of desire report at play behind this intuition is what I've called the **advisory** kind in Chapter 5. Attributers of such desire reports can help themselves to some of their information (of which the desirer herself may be ignorant) in working out what would put the desirer into preferred states, and attribute the desire on that basis. (Said of someone with no knowledge of the MTA: 'Sally is heading to Harlem? She doesn't know it, but she wants to take the A train'.) The paradox here is better understood as involving the more familiar **predictive** kind of desire attribution, because it should be no surprise that advisory uses are radically externalist.

[4] There's an ambiguity here in the phrase 'strongest desire' that hasn't mattered much until now. On one reading, it means a desire that is strongest in terms of *content*; that is, a desire whose content entails that of all the agent's other desires. On another reading, it means something like 'most psychologically salient/forceful desire', which may fail to be logically strongest. It's the former reading we need for these purposes.

nothing in the derivation of the contradiction depends very much on the notions of desire and satisfaction in particular. We could have formulated the paradox in terms of closely related attitudes and relations, such as *seeming* desire, or desires *seemingly* being satisfied, or seeming desires seemingly being satisfied, etc. The weaker these attitudes and relations get, the less plausible becomes the story of why the analogue of (SDSS) should fail; but the formal derivation of the paradox isn't affected thereby.[5]

(SDSS), applied to the Mal/Ben case, is all that we need to generate the biconditionals $m \leftrightarrow \neg b$ and $b \leftrightarrow m$. And classical logic is all that is needed to generate contradictions (and triviality) from these biconditionals.

I'll argue that certain principles of classical logic are indeed the root of the paradox, and that giving them up yields a *unified* solution to all paradoxes with this structure. I'll start making this case by drawing out analogies and disanalogies to the more traditional Liar paradox.

4.7 Non-classical thought

Readers familiar with semantic paradoxes will recall Kripke (1975)'s contingent paradoxes of self-reference. Mal's and Ben's desires are reminiscent of the following pair of sentences:

> The sentence just below this one isn't true.
>
> The sentence just above this one is true.

Is the upper sentence true, or not? If it's true, then the sentence below it isn't, which means that it can't be true either. Thus it must not be true. But then the sentence below it *is* true, in which case the upper sentence must be

[5] The situation here is similar to revenge versions of Prior's paradox discussed by Prior (1961), Bacon, Hawthorne, and Uzquiano (2016), Bacon and Uzquiano (2018), and Bacon (forthcoming). Our desire paradox, suitably formulated in a language with propositional quantifiers and a propositional definite description operator, can be seen as a contingent version of Prior's paradox. The contingency makes particularly vivid the costs of classical propositional logic in this context, for, unlike intrinsically paradoxical versions (in which, say, you believe that all of your beliefs are false or desire that all your desires be frustrated), there doesn't seem to be anything particularly irrational or unrealistic about Mal's and Ben's desires considered individually. See also Caie (2012) for similar paradoxes involving belief, where he argues that the rational response to paradoxes like the Liar is to adopt indeterminate states of belief, according to which you neither believe nor fail to believe the Liar sentence.

true after all. The reasoning that leads to paradox is similar to that at play in our desire paradox.

The *matter* of the paradoxes, however, is quite different. Self-referential sentences, like the Liar and Curry sentences, have been said to give rise to *semantic* paradoxes. It's commonly thought that these paradoxes stem, at bottom, from languages too expressive for their own good. Here, for example, is Ramsey explaining why we need to keep an object-language truth predicate at the cost of facing paradoxes:

> We get statements from which we cannot in ordinary language eliminate the words 'true' or 'false'. Thus if I say 'He is always right', I mean that the propositions he asserts are always true, and there does not seem to be any way of *expressing* this without using the word 'true'. (Ramsey 1927, my emphasis)

And here is Field explaining his view of the dialectic concerning paradoxes and classical logic:

> There is little reason to doubt the correctness of classical logic as applied to our most serious discourse, e.g. our most serious physical theories. But the semantic paradoxes arise because truth *talk* gives rise to some anomalous applications (e.g. "viciously self-referential" ones), and it's rash to assume that classical logic continues to be appropriate to these applications. (Field 2016, my emphasis)

Common to both of these passages is what I'll call a *linguistic* diagnosis of these paradoxes. Paradoxes arise, according to this diagnosis, because we want to express certain things (like infinite conjunctions—see Picollo and Schindler (2017)) that we cannot express in a finite way without a truth predicate that behaves disquotationally. This need for expressive power yields a tool that is, in some sense, too powerful for classical logic. The lesson is that a naïve truth predicate and classical connectives are two expressive tools that cannot be combined without triviality. It's because of the way we need to talk that paradoxes of self-reference arise.

The traditional menu of solutions developed in response is similarly informed by this linguistic conception of how the paradoxes get going. Their lesson, according to orthodoxy, is something like this: When developing a formal language, you have to be careful to avoid certain natural combinations of expressive tools that, when combined, yield triviality. Either you must avoid saddling a language with its own truth predicate (as Tarski and Russell urge). Or, if you do insist on having an object-language truth predicate, you've got to weaken it in some unnatural way—either by having sentences with semantic value 1 that aren't true, or by having sentences with semantic value 0 that are true. Or you've got to weaken the logic to something less powerful than classical logic. Perhaps Kleene's K_3; perhaps Priest's Logic of Paradox; perhaps something else. The point is, these paradoxes are thought to have something fundamentally to do with language and expressive power. Thus the label 'semantic'.

Our desire paradox, despite involving structurally similar reasoning, doesn't have anything to do with language in particular. It doesn't make use of any particularly semantic properties like truth or falsity. Instead, it's just about people getting what they want, or failing to. Of course, we *use* language to *talk* about people getting what they want; but when we do so, we aren't talking about language, as we're obviously doing when we attribute truth or falsity to sentences as in the traditional semantic paradoxes.

Indeed, the ability of Mal and Ben to speak any particular language is inessential to the paradox. Mal and Ben needn't have any thoughts about which sentences in which languages are true or false in order to desire what they desire. Each just needs to have attitudes about the other's desires, regardless of which language they are expressed in. Regardless, indeed, of whether they are expressed in *any* language.

To make the point particularly strongly, it may even be possible—although I do not stake much on this claim—for sufficiently sophisticated non-linguistic creatures to get themselves into desiderative situations similar to Mal's and Ben's. Desires are things that pretty much any sentient being can have. Horses desire to roam free, tigers to hunt their prey, antelope to flee their predators, all without speaking a language to communicate these desires. These non-linguistic creatures also have the capacity to form other-directed desires: A mother fox might desire the satisfaction of her offspring's desires, or a malicious cat the frustration of his owner's.[6] Thus we might as well imagine that Mal and Ben are two cats, who cannot speak but can be respectively malicious and benevolent. If that's right, it would make a particularly strong case that paradoxes with a Liar-like structure arise at the level of thought

[6] I have met such cats.

itself.

We know from the literature on contingent Liar sentences that the way to solve those paradoxes isn't to deny the existence of the problematic sentences. Paradoxical sentences can clearly be formulated in natural language, and a minimal amount of mathematics make them unavoidable in formal ones. I've argued above that this is so for paradoxical propositions as well: we should not deal with the problems they raise by denying that they exist. Instead we have to deal with them by modifying something else in our theories.

What kind of modification should that be? I'll conclude by suggesting that this paradox gives us a new abductive reason to prefer broadly non-classical treatments.

Solutions to the semantic paradoxes come in two flavours: those that keep classical logic, and those that revise some part of it. Those that keep classical logic must deny one of the T-Schemas:

$$T(\ulcorner s \urcorner) \rightarrow s \tag{T-out}$$

$$s \rightarrow T(\ulcorner s \urcorner) \tag{T-in}$$

The ability to retain these intuitive schemas unrestrictedly has long been touted (for example by Priest (1987) and Field (2008)) as a mark in favour of non-classical approaches. Most of the work that goes into classical approaches, therefore, is concerned with retaining something as close to these schemas as possible without reinviting paradox.

The non-logical (SDSS) is the clear analogue to the T-schema. Recall that this principle says:

> **Strongest desire satisfaction schema (SDSS)**: If *S* most strongly desires that *p*, then *S*'s strongest desire that *p* is satisfied if and only if *p*.

Defenders of classical logic will point to this principle as the culprit in the desire case, just as they point to the T-schema as the culprit in the Liar paradox.

However, the T-Schema and (SDSS) are about different things. The T-schema relates syntactic objects (sentences) to their truth conditions; (SDSS) relates desires to their satisfaction conditions. This makes a difference. For example, shifts in facts about meaning/conventions can change which sentences are true, but they cannot change what I desire (unless I desire something about meanings or conventions).

To see this, let us grant that ‘Beer is red’ is false, and that I currently desire red wine, though not beer. Consider the following counterfactuals:

1. If ‘beer’ had meant what ‘red wine’ means, ‘Beer is red’ would be true.
2. If ‘beer’ had meant what ‘red wine’ means, I would want a beer.

Plausibly, (1) is true, while (2) is false. Shifting the meaning of ‘beer’ changes the truth conditions of sentences, but not the satisfaction conditions of my wine-related desires. Thus, (SDSS) and the T-Schema are not just trivial notational variations. They have substantively different content, and different modal profiles. That means that anyone who wants to block the desire paradox by denying (SDSS) does not get that move for free, by retelling whatever story she told to deny the T-Schema. A different kind of story would have to be told in favour of giving up each one.

A non-classical approach, on the other hand, can solve the Liar paradox and the desire paradox by abandoning *the exact same principles*. Both paracomplete and paraconsistent approaches are possible. I’ll conclude by briefly sketching a paracomplete story along the lines of Field (2008), not because I think it’s inevitable, but because I favour it over paraconsistent approaches for a broad class of intensional semantic paradoxes (for example the Knower paradox—see Jerzak (2019) for a more involved technical exposition). I won’t delve into the technical details here; instead, I’ll show how it can yield an attractive package of results in the Mal/Ben case.

In a nutshell, a theory of this kind holds that it’s *indeterminate* who gets what they want in the Mal/Ben case. The claim that Mal gets what she wants falls into a truth-value gap, as does its negation. The claims that Mal gets what she wants if and only if Ben doesn’t, and that Ben gets what he wants if and only if Mal does, are true.[7] The classical argument from those biconditionals to explosion fails at the negation-introduction step: indeterminate claims are not such that we can suppose them, derive a contradiction, and infer their negations. When possibly indeterminate sentences/propositions are involved, we must be careful not to reason as if they have classical truth values.

[7] These biconditionals have to be formulated with a more complicated, non-truth functional conditional, instead of the material conditional. But this conditional collapses into the material conditional in bivalent contexts, so not much is lost by this.

This theory denies only two classical rules of inference, both involving suppositional reasoning: negation-introduction, and if-introduction.[8] A unifying virtue of this theory, in our context, is that it attributes the error in reasoning in the Liar paradox, and in the desire paradox, to the exact same steps—in this case, the negation-introduction step. Classical theories, on the other hand, must attribute the error to two different kinds of satisfaction/truth principles—indeed, both of which enjoy immense intuitive support. Thus the classical theorist sees disunity where we ought to have expected unity. The ability to solve two structurally similar paradoxes in the exact same move is a mark in favour of non-classical approaches.

Such a theory still comes with a certain kind of external supervenience—but only of the familiar kind involving truth value (rather than content). Facts about whether my desires are satisfied can have non-classical semantic values, and *that* will supervene on more than just my local situation. For instance, Mal's strongest desire would have a classical truth value if Ben's strongest desire were for a beer, and becomes indeterminate the moment Ben forms the desire ascribed above. However, we should sleep easier with this kind of externalism than with the radical externalism that the relational case-denier must espouse. It's a common idea that our beliefs and desires go from true to false, or from satisfied to unsatisfied, according to the whims of the world. On this paracomplete view, they can go from determinately satisfied/unsatisfied to indeterminate just as easily. However, facts about *what* I desire are more well-behaved. They may supervene widely, but only for the tractable, familiar reasons explored by Burge and Putnam.

This is a more attractive package of views, I think, than the classical alternatives. It all amounts to a new consideration in favour of non-classical approaches to paradoxes in the family of the Liar. Not only must classical approaches invalidate extremely plausible inferential principles involving truth (T-out and T-in), they must also either forbid certain seemingly possible combinations of desires, or else the independently plausible (SDSS). The non-classical approach I outlined solves both sentential and non-sentential kinds of paradoxes in the exact same move. This is a new mark in its favour.

However, the argument for this upshot relied on denying radical externalism. As we'll see in the next chapter, there are reasons to believe that desire reports in natural language can be radically externalist in the way I've denied here. It's to those arguments that I turn in the next chapter.

[8] Negation-introduction: $\varphi \vdash \bot \implies \vdash \neg\varphi$; if-introduction: $\varphi \vdash \psi \implies \vdash \varphi \rightarrow \psi$.

Chapter 5

Two Ways to Want?

5.0.1 In vino veritas

Too often have I suffered the misfortune of being directed to bring wine to a dinner party. Not beer, or whisky, both drinks whose quality I'd be quite a bit more competent to judge, but specifically wine. With the aid of my visual system I can usually distinguish the red stuff from the white stuff; that just about exhausts my ability to make discriminations.

What I want is the best wine for the occasion, which I understand to be that which will bring the most joy to my more gustatorily advanced dining comrades. I only care about my comrades' taste; all wines taste fine to me. There I stand, in the grocery store, having whittled the options down to two. There's a Zinfandel from Sonoma Valley, and a Sauvignon Blanc from New Zealand. Unbeknownst to me, the Zinfandel would bring my dinner companions the most joy; they find the Sauvignon Blanc's grassiness oppressive. You, a maximally informed observer of the situation, are looking at me in my predicament. A natural way for you to describe the situation—which I will refer to as *in vino veritas*—is with (1):

(1) He doesn't know it, but he wants the Zinfandel.

(2), however, rings false:

(2) #He doesn't know it, but he believes that the Zinfandel is the wine to get.

After all, if I believed that the Zinfandel were the wine to get, there would be no predicament—I'd simply get it, my comrades would savor it, and all would be well.

Suppose further that my dining companions change their taste, now finding the Sauvignon Blanc's grassiness pleasant and the Zinfandel's fruitiness overwhelming. (3) would then be the right thing to say:

(3) He doesn't know it, but he wants the Sauvignon Blanc.

Something strange has happened. Without changing anything about me—I've just been standing there dumbfounded all along—my desires seem to have changed. I went from wanting the Zinfandel to wanting the Sauvignon Blanc, without any corresponding change in my underlying psychological state.[1]

This contrasts with belief. Nothing about my beliefs has changed here. All along, I believe that whichever wine I buy should align with my comrade's preferences. What wine would make that true changes, but the content of my beliefs don't.[2] My beliefs about what's best to get are compatible with any situations in which the wine-to-get lines up with the wine-they-want. They in themselves mark no distinction between the Zinfandel and the Sauvignon Blanc. My desires, however, attributed in (1) and (3), seems to have a kind of sensitivity to the wider world that my beliefs do not. In situations where my comrades *actually* want the Zinfandel, whether I know it or not, there's a sense in which I want that too; and in situations where they actually want the Sauvignon Blanc, so, in some sense, do I.

You might resist these data.[3] Perhaps (1) and (3) aren't really true.

[1] This kind of case was brought to my attention by Callard (2017), although she is concerned there not to give a particularly realistic semantics for natural language desire attributions, but rather to argue on behalf of Socrates that we can never truly desire things that are (in fact) bad. This kind of use is also mentioned in Davis (1984), and in Rooryck (2017).

[2] Belief attributions sometimes bear *de re* readings, with behavior superficially similar to that of advisory desire attributions. If Susan has a general belief that all Minnesotans are nice, but no particular beliefs about some Minnesotan (Fred) whom she's never met or heard of, I could reasonably say, "Susan thinks Fred is nice". However, this phenomenon is more limited with 'believes' than with 'wants'. If Susan *had* met Fred, and, not knowing that he was Minnesotan, formed the definite opinion that there is nothing nice about Fred, such a *de re* belief report would be inappropriate. (1), on the other hand, is appropriate even if I'm erroneously convinced that my friends prefer the Sauvignon Blanc.

[3] Empirical work remains to determine how cross-linguistically robust these advisory uses are. Informal surveys suggest that it is harder, if not impossible, to hear in (for example)

Perhaps all I ever really wanted all along mirrored my beliefs about what's best to get; the content of my desire was, less determinately, to get some wine or other that pleases my comrades. Someone pressing this line would insist that the correct way for you, the better-informed observer, to describe the situation would be:

(4) He doesn't want the Zinfandel. (Not yet anyway. But he will once he learns that it's the wine that his dinner companions prefer.)

I don't find (4) a horribly unnatural thing to say. But it's no more natural than (1). And (1) and (4) seem inconsistent. This suggests that we're attributing desires in two different ways. In (1), information beyond my ken helps determine what I want. In (4), what I want more or less coincides with what I believe to be good.[4]

(1) and (4) are typical examples of two different uses we make of 'wants'. One use is to predict and explain how agents act, roughly along the lines of belief-desire folk psychology. If I know that someone wants *A*, and believes that doing *B* will result in her getting *A*, and nothing stands in her way of doing *B*, I'll usually predict that she'll do *B*. The use of 'wants' in (1) clearly isn't this notion; if all you're allowed to do is observe, not to advise, you won't predict that I'll toddle off to the party with the Zinfandel in hand. Indeed, if you knew that I falsely believed my dinner companions to prefer the Sauvignon Blanc, you'd make exactly the opposite prediction. Let's call this the **predictive** use. On the predictive use, (1) is false and (4) is true.

But there's definitely another sense in which, if I buy the Sauvignon Blanc, I won't have bought what I really wanted all along. Indeed, I'd readily admit as much once my error becomes known to me. Say that I, falsely believing

German, Spanish, and French. An anonymous reviewer helpfully pointed out that the 'wants' in English originally meant *lacks*, as in "the soup wants salt". Only later did the psychological use develop. It could be this evolutionary history that explains why English is unique here, if it turns out to be. The discussions of desire in Socratic dialogues like the *Meno*, *Gorgias*, and *Republic* suggest that such a reading was also available in ancient Greek, but I won't press this point. The interesting thing for our purposes is that there *is* an attitude verb in some language that exhibits the kind of information-sensitivity more commonly associated with modals.

[4] I say "more or less" in light of Lewis (1988)'s argument against such an identification. (Although see Bradley and Stefánsson (2016) for a counterargument.) The important point for these purposes, as will become clear in what follows, isn't the identification of desire with belief, but rather the identification of the *information* to which desire reports are sensitive with the desirer's own beliefs.

the Sauvignon Blanc to be preferred by my comrades, buy it and bring it to the party. It would be natural for me to express my regret with:

(5) Ach! That wasn't the wine that I wanted!

Or say that you, the maximally informed observer, break your silence to dispense advice. You'd say:

(6) Return that Sauvignon Blanc! That's not what you wanted, your comrades hate it. What you really wanted to buy was the Zinfandel.

It would be odd for me to retort:

(7) ?You're wrong! I really did want to buy the Sauvignon Blanc. I bought exactly what I wanted. But I've changed my mind, and *now* I want the Zinfandel.

It would be much more natural to retract my previous claim about what I desired, saying something like:

(8) Oh! You're right, I guess I didn't want the Sauvignon Blanc after all. Thanks for telling me.

Situations like these, where better-informed agents offer advice to worse-informed ones, are where we most often find the use of 'wants' that I'm interested in. We ask the subway worker which train we want to get on, given where we're going; a good sommelier *tells* you what wine you want, instead of sitting back and laughing at you while you select the Chardonnay you erroneously think will go nicely with your ribeye. This use of 'wants' isn't the predictive one. In telling you what you want, better-informed advisers like the subway-worker and the sommelier are making use of *their* information, not restricting themselves to yours. I'll call this the **advisory** use, since it figures most prominently in situations of advice.

In what follows, I take it as evident that we attribute desires in this advisory sense, and not just in fringe circumstances. Injunctions like, "Figure out what you really want, before you do anything you'll regret!" sound extremely natural, as do doubtful self-attributions, as in, "I think I want the 9am flight, but I won't know for sure until I know when the meeting is." Similar injunctions involving "believes" sound very weird. It's easier to be ignorant about what you want than about what you believe, and theories of attitude verbs shouldn't disallow that.

The plan is this. I rehearse two popular theories of desire attributions: Heim's restricted modal account and Levinson's decision-theoretic one, showing that both of them, being engineered with predictive uses in mind, don't predict advisory uses. I then present data concerning the interaction of desire reports with conditionals, where Heim's and Levinson's theories also founder. I develop a lexical ambiguity response, which posits a separate semantic entry for "wants" corresponding roughly to Socrates' apparent view: that we want what's good according to an *omniscient* state of information. I give two reasons for dissatisfaction with this response, and develop a better one according to which desire reports express information-neutral propositions. I compare "wants" to "ought", arguing that the former functions as a precisification of the latter. I take up the relationship between advisory and predictive uses, and develop a view according to which predictive uses, where they differ from advisory ones, are literally false; their apparent felicity is explained by free indirect discourse. If this view is right, then there are only apparently two different ways to want. Finally I sketch an account of the purpose of desire attributions that explains why it made sense for them to evolve this way.

5.1 Existing proposals

5.1.1 Warm-up: the naïve semantics

Here's a flat-footed first pass at modeling desire attributions. Agents have, at bottom, preferences concerning outcomes, and what they want is a function of those preferences. To say that an agent wants φ is just to say that the outcomes she most prefers are ones in which φ holds.

A well-known problem for this approach, discussed in Stalnaker (1984) p. 89, is that it predicts that I want whatever follows from, or is presupposed by, what I want. Say, for example, that John is sick, and would very much prefer not to be. It's true to say,

(9) John wants to get better.

But on the naïve semantics, this entails

(10) John wants now to be sick.

since every world in which John gets better is a world in which John is now sick. Therefore if “John gets better" is true throughout the worlds John considers best, “John is now sick” must also be true there. So if John wants to get better, he wants to be sick now. We'd expect him to protest this consequence, and our theory of desire attributions should not contradict him in this.

5.1.2 Stalnaker and Heim

This example shows that what we want isn't just a matter of what's going on in the worlds we most prefer. It also depends on which options are live in the situation we find ourselves in. John never wanted to be sick, but given that he is, he wants to get better. Thus what we want depends, in addition to basic preferences on outcomes, on a state of information—a state, that is, that includes certain options as live and rules out others as dead. It's our preferences regarding live options that factor into the truth conditions of a propositional desire report. Worlds in which John never got sick are not live options, so his preferences regarding them, strong though they might be, don't factor into characterizing his state of desire with respect to getting better.

How exactly does a state of information combine with basic preferences to yield desire attributions? A natural thought, first outlined by Stalnaker, is that I want φ if, throughout the live worlds in the relevant state of information, my basic preferences render nearby φ worlds better than nearby $\neg\varphi$ worlds. The question then becomes: which information is relevant? Hitherto the literature on desire attributions has implicitly assumed an answer to this question: The body of information that's relevant is that which characterizes the desirer's own beliefs. Stalnaker:

> Wanting something is preferring it to certain relevant alternatives, the relevant alternatives being those possibilities that the agent believes will be realized if he does not get what he wants. (Stalnaker (1984), 89)

Heim (1992), who fleshes out Stalnaker's idea formally, makes the same assumption. Some notation:

- $\succeq^w_x$: a preorder on worlds, so that $w_1 \succeq_{x,w} w_2$ just in case agent x in w weakly prefers w_1 to w_2 ($\succ^w_x$ for strong preference). For sets W_1, W_2 of worlds, $W_1 \succ^w_x W_2 := \forall w_1 \in W_1, \forall w_2 \in W_2, w_1 \succ^w_x w_2$;
- B^w_x: The set of worlds compatible with x's beliefs in w;
- $\text{Min}_w(\varphi)$: The set of most similar worlds to w in which φ holds.

With these resources in hand, Heim proposes the following semantics:[5]

$$[\![x \text{ wants } \varphi]\!]^w = 1 \text{ iff } \forall w' \in B^w_x, \mathit{Min}_{w'}(\varphi) \succ^w_x \ \mathit{Min}_{w'}(\neg\varphi).$$

Neither Stalnaker's idea nor Heim's formalization of it was engineered with cases like (1) in mind. This is easy to see just by sketching a model faithful to the structure of *in vino veritas* and showing that Heim's semantics does not churn out (1). An explicit model of the case and a derivation of Heim's truth conditions relative to it are sketched in the appendix.

Intuitively, though, it's easy to see why Heim's semantics doesn't produce (1). In the *in vino veritas* case, I have no beliefs about which wine my comrades prefer. Thus, while my basic preferences render worlds where my selection aligns with my comrades' tastes better than those where it doesn't, my beliefs do nothing to single out the Zinfandel. So it's not the case that, throughout all worlds compatible with my beliefs, nearby I-buy-the-Zinfandel worlds are preferred by me to nearby I-buy-the-Sauvignon Blanc worlds. There are counterexamples among the non-actual worlds, which my beliefs do not rule out, where my comrades prefer the Sauvignon Blanc. Thus Heim's semantics misses true readings in situations like *in vino veritas.*

5.1.3 Decision-theoretic accounts

Levinson (2003) also complains that Heim's semantics fails to validate intuitively true desire attributions. But his cases are quite different in spirit from mine, and motivate a different kind of theory from Heim's. Since I want a semantics that handles both kinds of cases (and, as we'll see in §III, combinations of them), it's instructive to consider his examples and

[5] Heim actually casts her proposal in the framework of dynamic semantics; I've reformulated her view in a static setting, since the dynamic framework is motivated by considerations, orthogonal to the present ones, about the projection of presuppositions in attitude ascriptions.

the decision-theoretic semantics he cooks up to accommodate them. I'll then show that his semantics doesn't help with the *in vino veritas* case, and consider ways to improve on it.

Levinson's case against Stalnaker and Heim involves insurance. Most of us, he observes, want to buy insurance sometimes. Even though it's pretty unlikely that our houses will burn down, it would be such a calamity if they did, that many of us want to be safe rather than sorry. But this poses a problem for Heim. For consider two worlds where my house doesn't burn down, but which differ as to whether I bought insurance. On the whole, do I prefer the one where I bought insurance, or the one where I didn't? I, for one, prefer the world where I hold onto my cash, instead of shelling out for an as-it-happens useless insurance policy.[6] But loads of the worlds consistent with my beliefs are worlds where my house won't burn down irrespective of whether I buy insurance. Therefore, I don't meet Heim's requirement that all of my belief-worlds render nearby "I buy insurance" worlds better than "I don't buy insurance" worlds.

To figure out whether someone wants to buy a particular insurance plan, we need information more fine-grained than anything on offer in Heim's semantics. Full beliefs and qualitative preferences aren't enough; we need to know just how likely she judges it to be that her house will burn down, how bad it would be for her if it did, and what the plan costs. These are quantitative, not qualitative matters.

Thankfully we have a quantitative theory of rational action at our disposal: Levinson's account avails itself of decision theory and its resources.[7] Let's upgrade Heim's less fine-grained ingredients accordingly:

- Upgrade $\succ_x^w$, a mere preorder on worlds, to an evaluation function

[6] Büring (2003) defends Heim against Levinson by arguing that those who buy insurance *do* prefer worlds in which they buy unused plans, because as long as they don't *know* that the plan will be useless, they primitively value the peace of mind that insurance brings in such worlds. This is, of course, a formal possibility; but Büring then owes us a new substantive account of preference, and I have trouble seeing how it could account for all insurance-style cases. There are gamblers and actuaries who make claim to make bets dispassionately, in the sense that they are perfectly psychologically at ease gambling and losing so long as the gamble was rational given their utilities and credences. That is, they explicitly claim not to primitively value peace of mind. It's hard to see how Büring could account for such cases, whatever substantive account of preference he gives. See Lassiter (2011) for further arguments in favor of a more fine-grained probabilistic framework.

[7] His account follows Goble (1996)'s decision-theoretic theory of deontic modals.

$g_x^w : W \to \mathbb{R}$, defined such that $g_x^w(w_1) \geq g_x^w(w_2)$ just in case agent x in w (weakly) prefers w_1 to w_2.

- Upgrade the state of information, previously identified with the set of worlds B_x^w, to what Yalcin (2012c) calls a **sharp information state** $i_x^w = \langle S_x^w, Pr_x^w \rangle$:[8]
  - ○ $S_x^w \subseteq W$ is the set of live epistemic possibilities for x in w;
  - ○ Pr_x^w : A probability function on W such that $Pr_x^w(S_x^w) = 1$.
- Shorthand: $Pr_x^w(w' \mid [\varphi])$: x's credence in w that w' is the actual world conditional on $[\varphi] = \{w : [\![\varphi]\!]^w = 1\}$.

Levinson proposes a semantics which says that you want φ just in case, relative to your credences and utilities, φ yields higher expected utility than $\neg\varphi$.[9] Formally,

$$\begin{aligned} [\![x \text{ wants } \varphi]\!]^w = 1 \quad & \text{iff} \quad EU_{x,w}(\varphi) > EU_{x,w}(\neg\varphi) \\ & \text{iff} \quad \sum_{w' \in S_x^w} g_x^w(w') Pr_x^w(w' \mid [\varphi]) \\ & \qquad > \sum_{w' \in S_x^w} g_x^w(w') Pr_x^w(w' \mid [\neg\varphi]). \end{aligned}$$

Levinson sketches an explicit model of the insurance case, and shows how his semantics predicts that 'you want to buy insurance' is true relative to it. Intuitively, while my full beliefs don't rule out that I'm in a situation where I shell out money for an as-it-happens useless plan, my quantitative preferences render an uninsured fire-ravaged house to be so calamitous an eventuality that, even though I judge it to be pretty unlikely, the calamitousness overwhelms the slim odds, making it worth shelling out a relatively small amount of money.

[8] Stipulate for simplicity that the set W of worlds is finite.

[9] This is a slight simplification of Levinson's official view. He actually defines 'wants' relative to evaluation functions g, in order to handle cases of active ambivalence between outcomes resulting in seemingly contradictory desire attributions. (E.g. "I want the wine [it will taste great], but I also don't want it [it will cause a hangover].") This is a different problem from the kind I'm interested in—in the *in vino veritas* case, what's going on isn't that you change how you feel about total outcomes, but rather that, given your fixed total preferences, different information states yield different results about what you want.

Thus Levinson predicts what Heim fails to predict—that I can want *p* even if not *all* of my belief worlds are ones where I prefer nearby *p* worlds to nearby not-*p* worlds. This is a virtue of his account. Plus, the decision-theoretic framework easily generalizes to graded desire attributions ("I *really* want beer"; "I want beer, but I want whisky way more"), whereas it's hard to see how Heim would have the tools for this.[10]

Does Levinson's semantics help with *in vino veritas*? Again, a quantitative model and derivation of truth conditions relative to a true-to-case model is sketched in the appendix. However, it's again easy to see intuitively why Levinson's semantics won't help. Just as my full beliefs and qualitative preferences don't change depending on my interlocutors' information, neither do my credences and utilities. Relative to them, I expect to be no better off buying the Zinfandel than buying the Sauvignon Blanc. Indeed, even if my credences and utilities rendered buying the Sauvignon Blanc the preferred action, the store adviser, having better information, can still felicitously correct my desire report. After he does this, I should retract any assertions to the effect that I wanted the Sauvignon Blanc. So Levinson, while improving on one aspect of Heim's proposal, does not solve our problem about advisory desire reports.[11]

5.2 'Wants' in the consequent of conditionals

The above shows that some true sentences involving 'wants' in certain contexts come out false on the theories offered by Heim and Levinson. In this section I'll show that their theories also fail to predict certain aspects of its compositional behavior. Back to the wine. Consider the following conditionals, said by you of me in the *in vino veritas* case:

(11) If his comrades prefer the Zinfandel, he wants the Zinfandel.

(12) If his comrades prefer the Sauvignon Blanc, he wants the Sauvignon Blanc.

[10] See Lassiter (2011) for a probabilistic account of modality that incorporates scales, familiar from the literature on gradable adjectives, to account for these data.

[11] Other proposals for the semantics of 'wants' exist: for example, those of von Fintel (1999), van Rooij (1999), Villalta (2000), Lassiter (2011), and Condoravdi and Lauer (2016). The differ in details, but all of them are fundamentally engineered to take the subject's doxastic state as the information relative to which the desire attribution is assessed.

These are both not only true, they are *extremely* true, in that they're among the most natural ways to describe the state of mind I'm in when I'm standing there dumbfounded in the store. Note here again the contrast with belief. (13) is extremely false:

(13) If his comrades prefer the Zinfandel, he believes that the Zinfandel is the wine to get.

You can use (11) and (12) to describe my conditional preferences, but (13) cannot be used to describe my conditional beliefs. (13) means that my beliefs are sensitive to my comrades' preferences, which, as a feature of the *in vino veritas* case, they are not. Granted, if I use a version of (13) first-personally, it doesn't sound *too* bad: "If my comrades prefer the Zinfandel, I think that's the wine to get."

But third personally it clearly doesn't work. To see this, consider a more knowledgable third party engaging in a bit of reasoning about what you want/believe. He would do ill to reason:

If his comrades prefer the Zinfandel, he believes that the Zinfandel is the wine to get.
His comrades prefer the Zinfandel.

He believes that the Zinfandel is the wine to get.

My comrades do prefer the Zinfandel, but I don't believe that the Zinfandel is the wine to get. The most plausible diagnosis of why this is bad reasoning is that the major premise is false; my beliefs aren't sensitive to my comrades' preferences, as it requires. However, given the availability of the advisory 'wants', he would do well to reason:

If his comrades prefer the Zinfandel, he wants the Zinfandel.
His comrades prefer the Zinfandel.

He wants the Zinfandel.

Indeed, this is exactly the kind of reasoning you'd engage in if wondering which bottle you should hand me.

This suggests that the maybe-vaguely-true-ish first-personal version of (13) is interpreted with the "thinks" taking wide scope over the conditional. This response is not available for (11) and (12), however. It would have it that sentences superficially of the form

$$\varphi \to x \text{ wants } \psi$$

are to be interpreted as

$$x \text{ wants } (\varphi \rightarrow \psi).$$

This approach has several shortcomings, of which I'll mention two. First, it doesn't validate the intuitively valid reasoning above, which results in your concluding that I want (in the advisory sense) the Zinfandel, as an instance of *modus ponens*. Perhaps the semantics of 'wants' could be fiddled with in such a way as to make $\{p, x \text{ wants } (p \rightarrow q)\}$ entail $\ulcorner x \text{ wants } q \urcorner$, but this also wouldn't be valid on Heim's or Levinson's semantics without modification. Since we'll need to modify the semantics anyway to make sense of the truth of these conditionals and the ability to reason with them using *modus ponens*, we might as well not butcher the surface grammar.

Second, the strategy crashes when the consequents are truth-functionally complex. Consider:

(14) If his comrades prefer the Zinfandel, then he wants to buy the Zinfandel, and (/but) they are snobs.

It's not clear how a defender of wide-scoping could interpret mixed conditionals like this. You might try:

$$me \text{ wants } (p_z \rightarrow (b_z \wedge snobs))$$

But this would be false—not wanting snobs for friends, but taking it to be quite possible that they prefer the Zinfandel, I certainly don't want it to be the case that, if my friends prefer the Zinfandel, they are snobs.[12] The best and simplest explanation here is that (11) and (12) are true, and have the logical form they seem to have.

Here's why Heim's and Levinson's accounts do not yield (11) and (12). I'll give a working semantics for the indicative conditional and show that (11) and (12) don't come out true in a moment. But first an informal gloss: A conditional is true in a context when, suppositionally adding the antecedent to the stock of information at that context, the consequent comes

[12] Something like this argument is present in Kolodny and MacFarlane (2010) for conditionals involving 'ought', and it traces back to Thomason (1981). The same mixed conditional would tell against an attempt to treat 'wants' as a primitive dyadic operator, of the form $\ulcorner x \text{ wants } (\varphi \mid \psi) \urcorner$. In general, the dialectic here mirrors the dialectic involving the interaction between deontic modals and conditionals. This, I argue in §V, is no accident, but illustrates deep structural similarities between 'wants' and 'ought'.

out true under that hypothesis. So add to the common information in a case like *in vino veritas* that my comrades prefer the Zinfandel. Is it true that I want the Zinfandel, according to Heim or Levinson? No—adding that information doesn't instruct us to change anything about my credences/beliefs or preferences/utilities. What I believe and prefer just depends on the world, not on the state of information in the common ground. So roughly speaking, the consequent will have the same truth conditions in the updated information state as in the non-updated one, and we've already seen that it's false with respect to those truth conditions in cases like *in vino veritas.*

Formally, I'll adopt a working semantics for $\rightarrow$ as a kind of epistemic modal.[13] On this view, a conditional functions as a test on the stock of information mutually presupposed in the conversational context: it tests whether adding the antecedent to that stock ensures that the consequent is true. Indicative conditionals are assessed relative to worlds and bodies of information i. Some definitions will be helpful. Shorthand: $[\varphi]_i = \{w \mid \llbracket\varphi\rrbracket^{w,i} = 1\}$.

Definition. An information state i **accepts** φ iff $[\varphi]_i = S_i$. In other words, iff $\forall w \in S_i, \llbracket\varphi\rrbracket^{w,i} = 1$.

Definition. The information state i **updated by** φ, written $i + \varphi$, is $\langle S_i \cap [\varphi]_i, Pr_i^\varphi \rangle$, where $Pr_i^\varphi(x) = Pr_i(x \mid [\varphi]_i)$.

The semantics for the indicative conditional $\rightarrow$ is then:

$$\llbracket\varphi \rightarrow \psi\rrbracket^{w,i} = 1 \quad \text{iff} \quad i + \varphi \text{ accepts } \psi.$$

It's a straightforward matter to verify (see appendix) that (11) and (12) both come out false on Levinson's semantics, combined with this conditional.

What kind of account might fare better? When we use conditionals like (11) and (12), we're describing something like my conditional preferences. Roughly speaking, (12) describes my state of mind when, restricting my attention to worlds in which my comrades prefer the Sauvignon Blanc, my preferences and *updated* credences judge I-buy-the-Sauvignon Blanc worlds to be better. To predict these truth conditions, we'll need the semantic value for 'wants' to be sensitive to the state of information that indicative conditionals operate on. That way, the antecedents of conditionals can modify the information parameter in the semantic entry for 'wants' in the right way.

[13] This kind of view is developed and defended in Yalcin (2007), Kolodny and MacFarlane (2010), and MacFarlane (2014).

This suggests that 'wants' belongs to the class of informational modals like epistemic might/must, deontic ought/may, and probability operators.[14] Indeed, I argue in §V, it functions as a systematic precisification of 'ought'.

I'll sketch two different proposals. The first posits a lexical ambiguity: a predictive entry governed by a semantics like Levinson's, and a "perfect information" entry which relativizes the information parameter to the state of perfect information at a world. I'll sketch some reasons for dissatisfaction with this bifurcation response, and then propose the semantics I'll ultimately endorse, according to which desire attributions express information neutral propositions.

5.3 Overreaction: perfect information

A natural reaction here would be twofold. First, since 'wants' does seem to have a sense, namely the predictive sense, more or less consonant with Levinson's semantics, one might posit a lexical ambiguity and use Levinson's semantics for 'wants$_{pred}$'. Second, one would add a new semantic entry for the advisory sense, 'wants$_{advise}$'. This semantics would have it that we want$_{advise}$ whatever our preferences judge to be better, not according to the state of information which characterizes our incomplete and possibly defective beliefs, but rather according to the state of perfect information. We really want what will *actually* put us into preferred worlds, in light of all facts known and unknown. That would suggest something like the following:

$$[\![x \text{ wants}_{advise}\ \varphi]\!]^{w} = 1 \text{ iff } \boldsymbol{Min}_{w}(\varphi) \succ_{x}^{w} \ \boldsymbol{Min}_{w}(\neg\varphi).$$

This semantics can predict the data of *in vino veritas*: relative to the actual world, nearby "I buy the Zinfandel" worlds are better according to me than nearby "I buy the Sauvignon Blanc" worlds. It can also predict the conditionals we've been interested in. Start out with a state of information that doesn't settle which wine my comrades prefer, and then update it with "my comrades prefer the Sauvignon Blanc." Relative to the worlds in this updated state, nearby worlds in which I buy the Sauvignon Blanc are better than those in which I buy the Zinfandel. So as far as the considerations on the table so far are concerned, bifurcation has everything going for it.

[14] Although see von Fintel (2012) for a defense of more classical approaches to these phenomena.

However, this response is an overreaction that we should reject for two reasons. First, it would make the advisory sense extremely difficult justifiably to use. Second, it can't account for true advisory uses in situations of known uncertainty—essentially, when Levinson-style insurance cases involve advisory aspects due to disagreement about the likelihoods of the relevant outcomes.

The first problem is simply that perfect information isn't easy to come by. To confidently assert that I want φ in the advisory sense, the perfect information semantics has it that you have to be fairly confident that, taking absolutely every consequence of my action throughout all time into account, I'll be better off by my own lights if φ than if $\neg\varphi$. That's quite a claim. Sure, you know that my comrades prefer the Zinfandel. But maybe but they are in such good spirits today that if I buy the Zinfandel, the party will be too rambunctious and we will all miss work tomorrow. Then I'd want$_{\textit{advise}}$ ***not*** to buy the Zinfandel. But maybe in addition to this all of our bosses will have taken the day off, and missing the day will have no immediate consequences. Then I'd want$_{\textit{advise}}$ the Zinfandel after all. But maybe, in addition to all of this, missing one day without consequence will instill in us a cavalier attitude towards punctuality, causing problems in our personal and professional lives. In this case, I don't want$_{\textit{advise}}$ the Zinfandel. And so on.

It might be claimed that this isn't so bad, since usually I can be reasonably confident, if never totally certain, that only relatively normal consequences will ensue from my comrades' enjoying a nice bottle of wine. So maybe we can never know for sure the truth of an advisory desire attribution, but we can often be justified in asserting them, and they can often turn out true.

But (the second problem) this simply gets the wrong result when I responsibly use the advisory 'wants' in cases where I don't have perfect information, and I'm perfectly aware that my advice probably conflicts with that of those with perfect information. Take a modified version of a Levinston-style insurance case:

> **Insurance-Arsonists**: You just declined to buy an insurance plan, because according to your credences and utilities, it was just barely too expensive to be worth it. However, I, unlike you, happen to know that a gang of arsonists has just moved to town. Thus the probability of your house burning down is much higher than you think it is—enough to tip the scales back in favor of

> your buying the plan. You've just finished telling the insurance salesman that you don't want the plan.

I speak truly when I say to you (in a whisper, naturally, so as not to tip off the lingering insurance salesman that his plan is probably mispriced):

(15) "No, that's wrong—you actually do want to buy this plan. I'll explain why later."

Now, as it happens, the gang of arsonists spares your house. So your house doesn't burn down, and you lose the money you spent on the plan. And, remember, you prefer no-housefire worlds in which you didn't shell out for the plan to ones where you did. Thus if wants$_{advise}$ is relativised to perfect information, I speak falsely in (15). This seems wrong. (15) seems like true and excellent advice, at least when I make it.

It's open to maintain that (15) *is* false, but to explain its seeming like good advice by holding that I was justified in asserting it. But it's hard to see why I would be justified in asserting it, if 'wants$_{advise}$' has these truth conditions. After all, when I assert (15), I know that it's still more probable than not that your house won't burn down, marauding arsonists notwithstanding. The arsonists aren't *that* efficient. Thus if the semantics of 'wants' in (15) were given by perfect information, I should think that (15) is very probably false when I assert it. So it's very difficult to see how I could nonetheless be justified in doing so.

The Insurance-Arsonists case suggests two things—one about the *source* of the information states that factor into the semantic values of advisory desire reports, the other about their *structure.* First, it suggests that the source of these information states isn't something that we can simply read off of the world of utterance. When advising people about what they really want, we aren't committing ourselves to something that only omniscient beings could know—that, taking account of absolutely every downstream consequence, you'll prefer the worlds that will/would result if the ascribed desire comes/came out true, compared to those in which it comes/came out false. The source of this information is more modest, and plausibly depends on context in some way.

Second, this case suggests that, whatever the *source* of these information states, their structure must be more fine-grained than that of Heim's semantics: they must represent some notion of likelihood, combined with a more fine-grained representation of preference. In the Insurance-Arsonists case, the metaphysically most similar worlds to ours in which you buy insurance are still worlds where your house does not burn. This is so even relative to the worlds doxastically accessible to me, the advisor. My information differs from yours not in terms of brute doxastic possibilities vis-a-vis house-burning: *both* of our doxastic possibilities include some housefire worlds and some no-housefire worlds, regardless of whether insurance is bought. In neither case will a semantics based on Heim's predict, even relative to the advisor's information, that you want to buy insurance. But this is wrong; my probabilistic information *can* make a difference to the truth value of a desire report. Thus whatever more flexible information base we relativize desire attributions to, that information base must include some representation of likelihood.[15]

One final reason to think that probabilistic structure is unavoidable, even on a more flexible account of the information source: desire ascriptions interact in non-trivial ways with probability operators in the antecedents of conditionals. Say that your roommate Ahmed, caring about your well-being and contemplating the possibility of rain, is advising you about whether to take an umbrella. There are two umbrellas in the house: a large and very effective one, and a small and moderately effective one. Your roommate is concerned about your not getting wet, but also about your not traipsing around unnecessary weight. We might communicate his desires concerning which umbrella you should take as follows:

(16) If it's probably not going to rain, Ahmed wants you not to take any umbrella.

(17) If it's probably going to rain, Ahmed wants you to take the small umbrella.

(18) If it's going to rain, Ahmed wants you to take the big umbrella.

[15] See Lassiter (2011) for further motivations for decision-theoretic semantics for a variety of modals. I do have some reservations about the standard EU approach here. For one thing it builds a huge amount of probabilistic and preferential coherence into the very meaning of desire reports in a way that seems implausible; see Buchak (2013) for discussion. What I take Insurance-Arsonists to show is that information states, even for advisory uses, must include some representation of likelihood. I've chosen the EU framework of Levinson because it's by far the most well-known account. There are less committal alternatives: see Holliday and Icard (2013) and Holliday, Icard, and Harrison-Trainor (2017). It's plausible that a more permissive theory would be more realistic, but delving into that more complicated machinery would unnecessarily cloud matters here.

Conditionals like these are easy to account for if the information states relative to which advisory desires are assessed have probabilistic structure. On a framework like Heim's, it's hard to see how such an account would go, since she only has qualitative doxastic possibilities in her toolbox.

5.4 Information-neutral desires

What, then, is the source of the information states that factor into the semantics of desire reports? It is not necessarily the desirer's: advisers can help themselves to information beyond that of the attributee herself. But it is not, as Socrates seems to have claimed, the omniscient information state. The information states that license even advisory desire attributions should still be human-sized, so to speak, and sensitive to probabilities and utilities in the way suggested by the Insurance-Arsonists case.

One could develop a contextualist semantics that indexes the information state to the attribut*er*'s information, but this is unpromising, for it wouldn't explain the genuine *disagreement* we seem to be in when we disagree about what someone really wants. If the proposition I express when *I* use the advisory 'wants' is indexed specifically to my information, and yours is specifically indexed to yours, then we simply talk past each other when we disagree. On the contextualist account, if you and a third party disagreed about which wine my comrades preferred, we should be happy to have the following exchange:

> You: "He wants the Zinfandel."
> Third party: "Well, yes, I agree, but he doesn't want the Zinfandel."

That should sound just as good as a long distance phone conversation running:

> You: "It's raining here."
> Third party: "Well, yes, I agree, but it's not raining here."

But it doesn't sound just as good. We're not talking past each other; we have genuinely incompatible views about what the agent really wants, not compatible views about what would put the agent in preferred states according to our respective information.

This dialectic is reminiscent of debates on epistemic and deontic modals. The most promising options for such information-sensitive vocabulary are some sort of flexible/group contextualism (Dowell (2011) and Dowell (2013)), expressivism (Yalcin (2012b)) and relativism (Kolodny and MacFarlane (2010)). For the sake of predictive concreteness, I'll sketch a relativistic version here, but my semantics can be easily adapted to expressivist or flexible contextualist background theories.

My proposal has two features. First, I'll model probabilistic informational common grounds with blunt probabilistic information states. Second, I introduce what I call 'mixed' expected utility functions $EU^i_{g^w_x}$, where the utilities come from one source (the agent x in world w), and the probabilities come from another (the blunt information states I representing the probabilistic common ground of the conversation). I'll explain these elements in turn.

The blunt information states relative to which semantic values of formulas are assigned are:[16]

Definition. A **blunt information state** I is a set of sharp information states $i = \langle S_i, Pr_i \rangle$, such that they agree on all the coarse-grained possibilities: $\forall i_1, i_2 \in I$, $S_1 = S_2$. (So it makes sense to speak of S_I.)

My proposal says that you want what yields highest expected utility according to your utilities, combined not with your credences, but instead with the probabilities of the information state in the common ground. First, a definition:

Definition. The **mixed expected utility** of φ, $EU^i_g(\varphi)$, relative to a utility function g and sharp information state $i = \langle S_i, Pr_i \rangle$, is the expected utility of φ derived from the probability function of i and the utility function g:

$$EU^i_g(\varphi) := \sum_{w' \in S_i} g(w') Pr_i(w' \mid [\varphi]_i)$$

My semantics uses these mixed functions. It goes:

[16] As Yalcin (2012c) shows, these kinds of states are in a much better position to represent probabilistic common grounds than sharp ones. They, unlike sharp states, can mark a difference between failing to render φ likely and positively rendering $\neg\varphi$ likely; for lots of propositions, our common information state doesn't say *anything* about their probabilities.

$$[\![x \text{ wants } \varphi]\!]^{w,I} = 1 \quad \text{iff} \quad \forall i \in I, EU^{i}_{g^{w}_{x}}(\varphi) > EU^{i}_{g^{w}_{x}}(\neg\varphi).$$

I will ultimately endorse this semantic entry, together with a relativistic postsemantics running as follows:[17]

> An utterance of the form $\ulcorner x$ wants $\varphi\urcorner$ is true as used at c_1 and assessed from c_2 if and only if $[\![x \text{ wants } \varphi]\!]^{w_{c_1}, I_{c_2}} = 1$.

This relativistic package, I'll argue, can predict the problematic data, and isn't saddled with the undesirable baggage of the bifurcation, perfect information response. I won't explain the entire relativistic semantic apparatus from the ground up—for that, see Bledin and MacFarlane. Instead, I'll walk through the predictions that this package uniquely makes. These predictions, I'll argue, are supported by the data, providing confirmation for this kind of approach.

5.4.1 Relativistic veritas in vino

According to this theory, desire attributions require two contexts to be assessed true or false: the context of use, and the context of assessment. So to judge the theory, we have to give a bit more information about who is asserting (1), and who is assessing it, in what kind of context.

Say that, in a context where I falsely believe that my comrades prefer the Sauvignon Blanc, I say to myself:

(19) I want the Sauvignon Blanc.

This is true as used and assessed relative to c_1, the context in which I utter it. This explains why I am justified in doing so. Now suppose that, later on in the shopping trip, you, having overheard (19), say:

(20) What you said before [in (19)] is actually wrong—you don't want the Sauvignon Blanc, you want the Zinfandel. That's the one your comrades prefer.

The relativistic semantics judges that, in this new context c_2, you are right; you've changed the context to include the information that my comrades

[17] See Bledin (2014) and especially MacFarlane (2014) for a general explanation of this relativistic framework.

prefer the Zinfandel. That means that (19), as used at c_1 and assessed at c_2, is false; relative to this better information, I want the Zinfandel, not the Sauvignon Blanc. Thus this package predicts that I'm obligated to retract (19), once I learn that my comrades prefer the Zinfandel. This is the correct result; the data of *in vino veritas* illustrate that it sounds very weird for me to stand by assertions like (19), once I acquire information relative to which my preferences render the opposite result. But it also predicts why it made sense for me to assert (19); assessed relative to the context of assertion, what I said was true.

It also yields, as the perfect-information semantics does not, the right results in the modified insurance case. When I learn about the marauding arsonists, my credence that nearby houses will burn rises. So when I whisper to you that you're wrong about wanting to decline the plan, I speak truly, relative to your context of utterance and my context of assessment. While *your* credences render the plan too expensive to be worth it, your utilities mixed with *my*, the assessor's, credences render the plan worth the money after all. I'm not asserting, falsely, that you *will* be better off buying the plan. I'm saying that it's the best option, relative to your utilities and what I know to be better information. That's why I give excellent advice when I tell you that you really want to buy the plan. Of course, if an even better-informed third party came along who knew that the arsonists planned to spare your house, then I should retract my assertion to the effect that you want the plan, and you should retract your retraction. This all jives extremely well with the information-neutral semantics, and isn't possible on the perfect information view.

5.4.2 Conditionals

Let's look at how the relativistic framework deals with the conditionals that were problematic for Heim and Levinson. Remember the conditionals:

(11) If his comrades prefer the Zinfandel, he wants the Zinfandel.

(12) If his comrades prefer the Sauvignon Blanc, he wants the Sauvignon Blanc.

The semantics for the indicative conditional carries over exactly from before, modified in a supervaluationist spirit to accommodate blunt probabilistic information states:

Definition. A blunt information state I accepts φ iff $\forall i \in I$, i accepts φ.

Definition. The blunt information state I **updated by** φ, written $I + \varphi$, is $\{\langle S_i \cap [\varphi]_I, Pr_i^{\varphi}\rangle \mid i \in I\}$, where $Pr_i^{\varphi}(x) = Pr_i(x \mid [\varphi]_I)$.

The semantics for the indicative conditional → is basically unchanged:

$$[\![\varphi \to \psi]\!]^{w,I} = 1 \quad \text{iff} \quad I + \varphi \text{ accepts } \psi.$$

The kind of information states where (11) and (12) are paradigmatically asserted are ones which include open worlds where my comrades prefer the Zinfandel, and open worlds where my comrades prefer the Sauvignon Blanc. I provide in the appendix a particular such information state, and show that the conditionals come out true. But again, intuitively, it's not hard to see what's going on. The antecedent of an indicative conditional like (11) restricts our attention to worlds in which my comrades prefer the Zinfandel, and asks what my expected utilities are, relative to information states including only *those* worlds, between my buying the Zinfandel and my buying the Sauvignon Blanc. Relative to these information states, my utilities render buying the Zinfandel the better option. So the indicative conditional is true relative to the original information state. *Mutatis mutandis* for (12).[18]

So, this relativistic semantics can predict the assertability/retraction data of the *in vino veritas* case. We've also seen that it, together with a plausible semantics for the indicative conditional, can predict conditionals like (11) and (12) in the contexts in which they seem true. Thus this relativistic theory has two predictive marks in its favor over previous ones, without falling prey to the inadequacies of the perfect information, bifurcation response.

[18] See the end of Appendix C for a consequence relation tracking information preservation on which modus ponens comes out valid. Interestingly modus tollens fails, and this is a good thing: In a context where we're ignorant about which wine my comrades desire, the following can plausibly all be true: A. If my comrades prefer the Zinfandel, I want the Zinfandel. B. It's not the case that I want the Zinfandel. C. My comrades prefer the Zinfandel. The situation is similar to that of Yalcin (2012a): the conditional is true in virtue of what *would* happen to the state of information after updating by the antecedent of the conditional; the desire attribution is false because relative to the original, more ignorant state of information, the Zinfandel and the Sauvignon Blanc yield equal expected utility; and the statement of my comrades' actual preferences is just a plain fact about the world.

5.5 'Wants' and 'Ought'

On the view I've offered, desire attributions function not only to predict what agents will do, but also to advise them about what courses of action they should undertake, if they want to realize their aims. To assert that someone wants φ is to communicate that, relative to her preferences and the best information available, she'll be better off by her own lights bringing about φ rather than $\neg\varphi$. That's not far from what we sometimes communicate with 'ought'. Telling an agent what she really wants is basically a way of telling her what she ought to do, given her basic aims, but our information about how to achieve those aims.

This similarity is unsurprising, for 'wants' and 'ought' pattern similarly. Just a few examples:

- Ross' puzzle:
  - x wants $\varphi \nvDash x$ wants $(\varphi \vee \psi)$;
  - x ought to $\varphi \nvDash x$ ought to $(\varphi \vee \psi)$.[19]
- Puzzling assertion/retraction data:
  - Both $\ulcorner x$ wants $\varphi\urcorner$ and $\ulcorner x$ ought to $\varphi\urcorner$ sound fine to assert if, relative to the common information at the context of assertion, x can expect to be better off by her own lights supposing φ than supposing $\neg\varphi$; but such assertions must be retracted if new information comes to light under which the opposite holds.
- Puzzling interaction with conditionals:
  - Both $\ulcorner\varphi \rightarrow x$ wants $\psi\urcorner$ and $\ulcorner\varphi \rightarrow x$ ought to $\psi\urcorner$ can be used to express conditional obligations/desires, motivating views of the indicative conditional as a kind of modal restrictor.[20]

On my view, 'wants' is a precisification of 'ought'—one which clarifies the *kind* of advice that is being given to agents. 'Ought' has notoriously many senses. If I claim that you ought to φ, I could be trying to communicate

[19] Ross' puzzle is solved on the decision-theoretic semantics. A state of information and utility function could give φ higher expected utility than $\neg\varphi$, while failing to give higher expected utility to $\varphi \vee \psi$ than $\neg(\varphi \vee \psi)$.

[20] See Kolodny and MacFarlane (2010), Yalcin (2012a), and Bledin (2014).

one of at least three things. I could be communicating that the better thing for you to bring about, given your subjective preferences and your subjective information, is φ rather than $\neg\varphi$. (“Oh well—even though your gamble didn’t pay off and the prize you would have won is horribly ugly, you did what you ought to have done.") Or I could be communicating that, relative to your preferences, but *my* information, your basic ends are more likely to be achieved by bringing about φ rather than bringing about $\neg\varphi$. (“Stop, you ought not buy the Sauvignon Blanc! Even though *I* hate it and think that everyone who prefers it is a snob, your comrades will be much happier with the Zinfandel, and that’s what you care about.") Or I could be communicating my disagreement with your ends themselves, represented by your preferences on worlds. (“You ought to buy the Sauvignon Blanc, even though your comrades hate it! Your comrades are snobs.") My theory of desire attributions predicts that only the first two of these three meanings is available for ‘wants’.[21]

This prediction is supported by data about how ‘wants’ and ‘ought’ embed differently under other attitude verbs. I’ll focus here on ‘thinks’. Consider Fred, a fellow dining comrade in the *in vino veritas* case. Fred knows that my comrades prefer the Zinfandel. He alone prefers the Sauvignon Blanc, and furthermore he is a solipsistic hedonist; he thinks that only his preferences should be taken into account when people are deciding what to do. My basic preference is to please as many of my comrades as possible, without any special provision for Fred. What does Fred think about all of this? I could describe Fred’s attitudes as follows:

(21) Fred thinks that, although *I* think I want to buy the Sauvignon Blanc, I actually want to buy the Zinfandel.

After all, he knows my preference is to please the majority of my comrades, and he knows that my comrades prefer it. However, it doesn’t seem right to say:

(22) ?Fred thinks that, although *I* think I ought to buy the Sauvignon

[21] Schroeder (2011) also contrasts ‘wants’ with ‘ought’, but the differences he highlights are orthogonal to those that I’m interested in. He points out that ‘wants’ functions as a control verb, while ‘ought’ is ambiguous between a control verb (which builds in an agent, as in “John ought to ski") and a raising verb (which operates solely on propositions, as in, “it ought to be the case that John skis"). The differences I highlight arise as a distinction between ‘wants’ and the control verb sense of ‘ought’.

Blanc, I actually ought to buy the Zinfandel.

If Fred thinks that I ought to buy the Zinfandel, then Fred *himself* prefers that I buy the Zinfandel. But Fred doesn't prefer this; he's a solipsistic hedonist, and only cares about getting his treasured Sauvignon Blanc. He thinks I *ought* to buy the Sauvignon Blanc, even though what I *really* want is to buy the Zinfandel, even though what I *think* I want is to buy the Sauvignon Blanc.

That suggests that the point of having an advisory 'wants' is to have a linguistic device that behaves like 'ought' with respect to information, but which rigidly fixes the agent whose preferences we're evaluating the relevant possibilities with respect to. A claim using 'ought' leaves undetermined whether I'm adding to the common ground my own information, or my own preferences, or both; a claim involving the advisory 'wants' clarifies that I'm only concerned with the information component. Thus the advisory 'wants' clarifies the *kind* of advice I'm giving the agent. Whereas 'ought' can give moral advice about how the agents' preferences should ideally go, 'wants' can only give pragmatic advice about how agents can best achieve their given aims.

5.6 Whither the predictive 'wants'?

I haven't said much about the predictive sense of 'wants', the only sense hitherto accounted for in the literature. What's the relation between predictive uses and advisory uses?

The first thing to point out is that my modification to Levinson's semantics is, in many ways, pretty conservative. Usually—not always, but usually—agents are aware of what consequences various actions are likely to bring about. In a large number of central cases, the attributee of a desire attribution is in more or less the same state of information as the attributers. Thus predictive and advisory uses can be expected to coincide in tons of cases. I attribute to you a desire to have one of the beers in the fridge; rarely do I have unique access to evidence that the beer is poisoned, or that the refrigerator is full of malevolent hobgoblins whom it would be better to leave undisturbed. This explains in large part why the advisory uses illustrated by cases like *in vino veritas* have gone unnoticed until now. Stalnaker, Heim, and Levinson focused on cases where there's no interesting asymmetry in information regarding the likely consequences of the desire's content.

Nonetheless, in cases where there *is* such an asymmetry, advisory and predictive uses come apart. So we need to tell some story about the also fine-sounding but incompatible predictive uses. I offer two possibilities, one more radical than the other. The non-radical proposal posits ambiguity; the more radical proposal attempts to account for predictive uses with only the advisory semantic entry, together with general principles concerning assertion. Ultimately, I suggest, the choice between them comes down to empirical questions about the cross-linguistic robustness of advisory uses.

5.6.1 Lexical ambiguity

The ambiguity view is exactly what it sounds like, and doesn't need much explanation. According to it, we simply have two semantic entries for 'wants': one where both the preferences and the information are hardwired to those of the desirer (Levinson's semantics), and one where the preferences are rigidly indexed to the desirer but the information state is variable. We use one 'wants' to predict what agents will do (in this sense I don't want the Zinfandel) and one to advise them about how to best satisfy their preferences (in this sense I do want the Zinfandel), given the information that's live in the relevant context. This is the view I'd fall back on, if the non-ambiguity view sketched below proves unworkable.

5.6.2 Non-ambiguity

The ambiguity view posits two semantic entries. It seems, at first glance, unavoidable to say something like this. After all, aren't (1) and (4) both true, in different senses, in the *in vino veritas* case, when both uttered and assessed in the same contexts?

Maybe not. It's possible to explain predictive uses, where they differ from advisory ones, with a single advisory entry together with general principles allowing us sometimes to take up the agent's perspective in making assertions. It's not so uncommon an idea that, when we're engaged in the project of explaining and predicting the behavior of agents, we sometimes utter sentences we know to be false, by way of describing the world as it looks from the agent's perspective.[22] Some examples:

[22] See Schlenker (2004) for background.

- (One police detective to the other, having previously taken the treasure out of the thief's hiding spot): A: "Why is the thief furiously digging there?" B: "He *knows* that the treasure is buried there."
- (In a context where we all know that Achilles hasn't defected to Athens): A: "Why haven't the Trojans invaded Athens yet?" B: "Achilles might have defected to Athens."
- (Said among fellow infidels:) A: "Why is that guy reciting the Athenesian Creed every morning?" B: "If he doesn't, God will smite him."

In none of these cases do we want to use the explanatoriness of the explanans as evidence for fiddling with the semantic entries of their components. In the first case that would give us non-factive knowledge; in the second, a semantics of "might" on which "might *p*" is compatible with "not *p*"; in the last, a theory on which it's fine for atheists to say that God exists and occasionally smites people.

In cases like these, I can successfully explain why someone did something, or predict that they are about to do something, by uttering sentences which I know to be false in my context. I know that it's not true that thief knows that the treasure is buried there. I utter that sentence by way of describing what the thief takes the world to be like, not what the world is actually like. Same with the other two cases: I know that it's not true that Achilles might have defected, because I know that he didn't defect; but I say that anyway, sketching the world as the Trojans conceive of it, to explain why they're not sending their legions. And describing the world according to the God-fearing man, as if it were actual, can explain why he's muttering the Athenesian Creed each morning.

The non-ambiguity view of the predictive 'wants' holds that the same phenomenon occurs when we use 'wants' to predict and explain agents' actions, in cases where we know that performing that action won't likely satisfy the agent's preferences. In *in vino veritas*, not only *is* there a reading on which (1) is true and (4) is false; that's the only reading that is literally true. There's no sense at all in which I want the Sauvignon Blanc, even if I'm doing everything in my power to buy it, because it's not what will actually satisfy my preferences relative to the information available to those asserting (1). But they can still talk as if I wanted it when they are predicting what I will leave the store with, because I *take* myself to want it. Thus the fine-sounding explanation:

- (Conversation between you and a bystander who also knows that my comrades prefer the Zinfandel): Bystander: “Why is that guy reaching up to that high shelf?" You: “Because that’s where the Sauvignon Blanc is, and he wants the Sauvignon Blanc."

On the non-ambiguity view, you just explained my action using a sentence you know to be false in your context. There’s not a different entry for ‘wants’ that tracks what agents *believe* will satisfy their preferences; instead, there’s a different kind of speech act that licenses unembedded quasi-assertions of false sentences, the believing of which makes sense of an agent’s behavior.

Is this plausible? To assess this, we'd need a good theory of this general phenomenon against which to measure the data we find for ‘wants’. Here are two relatively flat-footed considerations in its favor. First, it avoids lexical ambiguity, which is always nice when possible. Second, the predictive ‘wants’ patterns in some key ways like the other dialogues above. One feature paradigmatic of such explanations is that you can coherently continue the dialogue by asserting the negation of the just-seemingly-asserted explanans. In the treasure case, you can coherently continue: “Of course, the thief doesn’t *really* know that the treasure is buried there, because it’s in our police car." In the Achilles case, you can coherently continue, “Of course, it’s not really the case that Achilles might have defected; we all know he didn’t." In the God case: “Of course, that’s ridiculous; there’s no God, and even if there were he wouldn’t smite you for forgetting to recite an occasional Athenesian Creed." And—maybe—in the *in vino veritas* case: “Of course, he doesn’t *really* want the Sauvignon Blanc, because his comrades prefer the Zinfandel. He really wants the Zinfandel, and someone should go tell him that."

There are, however, some considerations against non-ambiguity. Predictive uses of ‘wants’ are very common, especially cross-linguistically (see footnote 3). If it turns out that English is unique in containing advisory uses, that would lend credence to the idea that it has some special word for expressing it. On the other hand, if we find that other languages sometimes contain desire reports whose relevant states of information don’t necessarily coincide with the desirer’s, that would support a non-ambiguity theory, on which the information state is variable at the level of the semantics. So the choice between these two options may depend on these empirical matters.[23]

[23] Thanks to an anonymous reviewer for pressing this point. Rooryck (2017) points out that, even in English, advisory uses are very hard to hear in the first person. Ambiguity

5.7 The purpose of desire attributions

It's one thing to give a relativistic semantics for 'wants' that makes some good predictions in cases that make trouble for other semantics. It's another thing to give some deeper explanation for why a natural language might have developed such a tool. Can it really be that what I want isn't just a function of what the world is like, but also depends on who is attributing the desire to me, and what information *they* have? I want to conclude here with a brief pragmatic sketch of why 'wants' might have evolved in this way.[24]

What is it to attribute a desire to somebody? One clear answer is the predictive one that I mostly haven't been concerned with here: it's to claim, of that person, that they are psychologically motivated to make the content of that desire come true. If this were the only use we had for attributing desires, we would never utter sentences like (1) in contexts like *in vino veritas*.

But our desires are not all of a piece. We want some things in virtue of wanting other things. I never just want to get on a particular train, end of story; I want to get on that train because it's the train going to Berlin, and I want to go to Berlin. And I don't just want to go to Berlin, either; I want to go to Berlin because that's where my friend is having her birthday party, and I want to be there to help her lament the passing of the years. Plausibly, these chains of explanation eventually bottom out; some things I just *want*, like (maybe) pleasure, or the Good.

The fact that our desires have this kind of structure opens up space for the possibility that you, knowing the general structure of my desires, have access to facts that interfere with these chains of dependence, facts that I myself don't know. Maybe this particular train, which I think I want to get on in order to go to Berlin, *isn't* the train to Berlin, but rather the mislabeled train to Paris. There you are, in the train station bidding me adieu, generally aware of the structure of my desires, just having noticed that I'm about to step on the wrong train.

In a case like this, it makes sense to have a linguistic device to communicate that my preference to go to Berlin stands a much greater chance of

theories can explain this by positing a difference in the two semantic entries for 'wants'. Non-ambiguity theories can account for this too: when I assert, in the present tense, that *I* want the Sauvignon Blanc, the information state parameter is saturated with my information in that context. So the non-ambiguity theory predicts that only a predictive reading is available in such cases.

[24] This kind of strategy is in the spirit of MacFarlane (2014), ch. 12).

being satisfied if I *don't* get on the train I think I want to get on. How are you to get this across? You *could* say, “You ought not get on that train!", but I might misconstrue what you mean. Maybe you've been insisting all along that Berlin is a den of Sin and Debauchery, and have been arguing the whole time that I ought not go to Berlin (even though you are fully aware that, relative to my rather more hedonistic preferences, Sin and Debauchery are things to seek out, not to avoid). What you need is a linguistic device to communicate that you A) are generally aware what my preferences are, and B) have information relative to which they'll actually stand a better chance of being satisfied if I do something other than what I think I want to do. English might have developed any number of such devices, but the one that actually developed is the advisory ‘wants’. You yell: “Stop! You don't want to get on that train!" and thereby accomplish exactly your communicative aim.

It's due to this function that ‘wants’ came to be assessment-sensitive. On the view I've offered, to add a claim of the form $\ulcorner x$ wants $\varphi\urcorner$ to the common ground of a conversation is to assert that φ outcomes are better than $\neg\varphi$ outcomes, relative to x's base preferences and our best information. Insofar as we care about x's preferences being satisfied, we should strive to make φ, and not $\neg\varphi$, come true. And if we subsequently acquire more information, information according to which x's preferences now render $\neg\varphi$ better than φ, we're obliged to take back our prior assertion. To stand by it is to let linger false information about what would be good for x by her lights. That is why it made sense for ‘wants’ to evolve to be assessment-sensitive; keeping a tally of who wants what is a way of keeping track of what should be done, if we want to help people realize their aims.

Appendix C

Consider the following language:

$$t := n_i$$
$$At := p_i$$
$$\varphi ::= At \mid \neg\varphi \mid (\varphi \vee \varphi) \mid (\varphi \wedge \varphi) \mid (\varphi \rightarrow \varphi) \mid t \text{ wants } \varphi$$

Let's think of t as a set of names of agents, and At as a set of propositional atoms.

Some abbreviations for readability: let $me = n_1$ with the intended interpretation of me, let $p_z = p_1$ with the intended interpretation of 'my friends prefer the Zinfandel', $p_s = p_2$ for 'my friends prefer the Sauvignon Blanc', $b_z = p_3$ for 'I buy the Zinfandel', and $b_s = p_4$ for 'I buy the Sauvignon Blanc'.

A base model $\mathcal{M}$ for the 'wants'-free fragment of this little language is a pair $\langle W, v\rangle$. W is a finite, non-empty set of worlds, and v is an interpretation function that sends every atom-world pair $\langle w, p\rangle$ to $\{0, 1\}$. Semantic values of formulas are defined relative to models, worlds, and blunt information states. Some definitions:

Definition. A **sharp information state** i relative to $\mathcal{M}$ is a pair $\langle S_i, Pr_i\rangle$, where $S_i \subseteq W$ and Pr_i is a function $\mathcal{A} \to \mathbb{R}_{[0,1]}$, for $\mathcal{A}$ a Boolean algebra of subsets of W, such that $Pr_i(S_i) = 1$, and for disjoint $A, B \in \mathcal{A}$, $Pr_i(A \cup B) = Pr_i(A) + Pr_i(B)$.

Definition. A **blunt information state** I is a set of sharp information states i such that $\forall i, i' \in I, S_i = S_{i'}$. (Thus we speak without ambiguity of S_I.)

Definition. $[\varphi]_I = \{w \in S_I : [\![\varphi]\!]^{\mathcal{M},w,I} = 1\}$.

Definition. The blunt information state I **updated by** φ, written $I + \varphi$, is $\{\langle S_i \cap [\varphi]_I, Pr_i^{\varphi}\rangle \mid i \in I\}$, where $Pr_i^{\varphi}(x) = Pr_i(x \mid [\varphi]_I)$.[25]

Definition. $[\varphi] = \{w \in W : \forall I, [\![\varphi]\!]^{\mathcal{M},w,I} = 1\}$.

Definition. A blunt state of information I **accepts** φ in $\mathcal{M}$ just in case, for all $w \in S_I$, $[\![\varphi]\!]^{\mathcal{M},w,I}$ = 1. In other words, if $[\varphi]_I = I$.

Semantic values of formulas are defined relative to models, worlds, and blunt information states:

$$[\![p]\!]^{\mathcal{M},w,I} = 1 \text{ iff } v(w,p) = 1;$$
$$[\![\varphi \wedge \psi]\!]^{\mathcal{M},w,I} = 1 \text{ iff } [\![\varphi]\!]^{\mathcal{M},w,I} = 1 \text{ and } [\![\psi]\!]^{\mathcal{M},w,I} = 1;$$
$$[\![\varphi \vee \psi]\!]^{\mathcal{M},w,I} = 1 \text{ iff } [\![\varphi]\!]^{\mathcal{M},w,I} = 1 \text{ or } [\![\psi]\!]^{\mathcal{M},w,I} = 1;$$

[25] $I + \varphi$ is undefined if $[\varphi]_I$ is empty, which leads to some counterintuitive results and some nice results. One of the counterintuitive ones is that my semantics predicts that you can't want what it's absolutely informationally certain you won't do. But I'm not so concerned with those results here; in paradigmatic instances of the advisory use, namely a context of advice-giving, you think it's not impossible that your advice will be heeded. Causal decision theory, which builds counterfactual notions into the definition of conditional probability, could help here.

$$[\![\neg\varphi]\!]^{\mathcal{M},w,I} = 1 \text{ iff } [\![\varphi]\!]^{\mathcal{M},w,I} = 0;$$
$$[\![\varphi \to \psi]\!]^{\mathcal{M},w,I} = 1 \text{ iff } I + \varphi \text{ accepts } \psi.$$

Heim

A Heim model $\mathcal{M}$ is a tuple $\langle W, Ag, Min, \succeq, B, v\rangle$. W is as before a finite, non-empty set of worlds. Ag is a set of agents. v, in addition to assigning semantic values to atoms, assigns members of Ag to agent names t. Min assigns to each world w a selection function $Min_w : \mathcal{P}(W) \to \mathcal{P}(W)$, where $Min_w(A)$ is the subset of A most similar to w. $\succeq$ assigns, for each world w and agent α, a preorder $\succeq^w_\alpha$ on worlds, representing α's preferences on outcomes w. B assigns, for each agent α and world w, a set B^w_α of of worlds compatible with the beliefs of α in w. As a convention, in models where $\succeq^w_\alpha$ and/or B^w_α do not depend on w, we write simply $\succeq_\alpha$ and/or B_α, respectively. For sets W_1, W_2 of worlds, $W_1 \succ^w_\alpha W_2 := \forall w_1 \in W_1, \forall w_2 \in W_2, w_1 \succ^w_\alpha w_2$. With these models, Heim's semantics runs:

$$[\![t \text{ wants } \varphi]\!]^{\mathcal{M},w,I} = 1 \text{ iff } \forall w' \in B^w_{v(t)}, Min_{w'}([\varphi]) \succ^w_{v(t)} Min_{w'}([\neg\varphi])$$

Here's a relatively realistic model of ***in vino veritas***. Let's make it a sad model, in which, in the actual world w_3, I buy the Sauvignon Blanc, but my friends prefer the Zinfandel. $W = \{w_1, w_2, w_3, w_4\}$, and $v(me) = $ me. Also Min is strongly centered in the model: every φ world is its own unique closest φ world.

The labeled solid lines represent Min_w; they point from a world to its closest neighbor(s) in which the label is true. (So the metaphysically most similar world to w_1 in which I buy the Zinfandel instead of the Sauvignon Blanc is w_2—after all, my decision about which to buy won't affect my comrades' taste.) No basic belief or preference change is included in the model, so we speak of B_{me} and $\succ_{\text{me}}$. B_{me} is the entire set—we're thinking of a time before I've made any decisions, so all possibilities are open. Thus at all worlds, all four possibilities are live options in w, in the sense important for Heim's semantics: my beliefs don't (yet) rule any of them out.

(1) teaches us that, in a situation like this, $\ulcorner me$ wants $b_z\urcorner$ should have a true reading. And Heim's semantics does not give us this. This is easy to see:

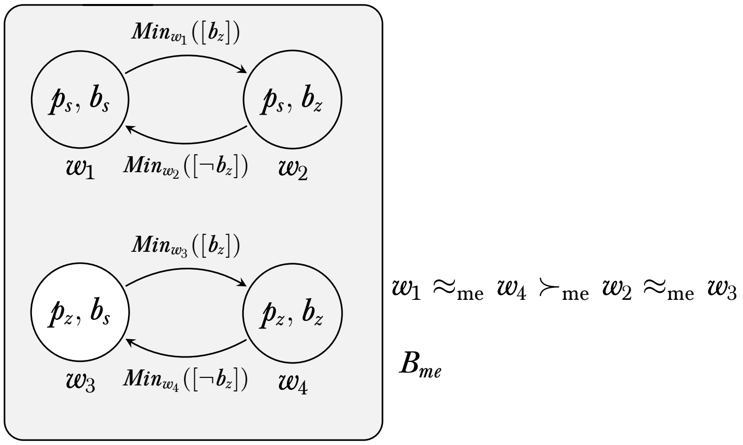

Figure 5.1: A model of *in vino veritas*

$$\begin{aligned}
[\![me \text{ wants } b_z]\!]^{\mathcal{M}, w_3, I} = 1 \quad & \text{iff} \quad \forall w' \in B_{v(me)}, Min_{w'}([b_z]) \succ_{v(me)} Min_{w'}([\neg b_z]) \\
& \text{only if} \quad Min_{w_1}([b_z]) \succ_{\text{me}} Min_{w_1}([\neg b_z]) \\
& \text{only if} \quad \{w_2\} \succ_{\text{me}} \{w_1\} \\
& \text{only if} \quad w_2 \succ_{\text{me}} w_1 \\
& \text{only if} \quad \bot.
\end{aligned}$$

Heim's semantic entry for 'wants' is not information-sensitive, so it doesn't matter what I is; for any I, w_1 is a counter-instance to the universal quantifier. Therefore Heim predicts that $\ulcorner me$ wants $b_z \urcorner$ is false in w_3. This is why Heim's semantics fails in predicting the advisory use.

Her semantics also doesn't predict the conditionals (11) and (12). These conditionals, in our language, are:

(11) $p_z \rightarrow me$ wants b_z

(12) $p_s \rightarrow me$ wants b_s

The paradigmatic kinds of information states where conditionals like these are asserted are those with open possibilities in which my comrades prefer the Zinfandel, and open possibilities in which my comrades prefer the Sauvignon Blanc. Thus let $I = \{\langle \{w_1, w_2, w_3, w_4\}, Pr_i \rangle\}$ such that $Pr_i(w) = .25$ for all $w \in S_I$.

$$
\begin{aligned}
[\![p_z \rightarrow \textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I} = 1 \quad & \text{iff} \quad \forall w' \in S_{I+p_z}, [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w',I+p_z} = 1 \\
& \text{iff} \quad [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I+p_z} = 1 \\
& \qquad \text{and } [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_4,I+p_z} = 1 \\
& \text{only if} \quad [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I+p_z} = 1 \\
& \text{iff} \quad \bot. \qquad \text{(Same calculation as above.)}
\end{aligned}
$$

Mutatis mutandis for (12). Thus Heim doesn't predict these conditionals relative to natural models of them, and information states relative to which they are naturally asserted, using a pretty natural semantics for →.

Levinson

A Levinson model $\mathcal{M}$ is a tuple $\langle W, Ag, g, Cr, v\rangle$. W is as before a finite, non-empty set of worlds. Ag is a set of agents. v, in addition to assigning semantic values to atoms, assigns members of Ag to agent names t. g assigns, to each $\alpha \in Ag$ and $w \in W$, a utility function $g^w_\alpha : W \rightarrow \mathbb{R}$. Cr assigns, to each $\alpha \in Ag$ and $w \in W$, a sharp information state $Cr^w_\alpha = \langle S^w_\alpha, Pr^w_\alpha\rangle$, representing that agent's epistemic possibilities and credences. As before, we conventionally drop the world superscripts for g^w_α and Cr^w_α, in models where these are stable across worlds. With these models, Levinson's semantics runs:

$$
\begin{aligned}
[\![x \text{ wants } \varphi]\!]^{\mathcal{M},w,I} = 1 \quad & \text{iff} \quad EU_{x,w}([\varphi]) > EU_{x,w}([\neg\varphi]) \\
& \text{iff} \quad \sum_{w' \in S^w_{v(x)}} g^w_{v(x)}(w') Pr^w_{v(x)}(w' \mid [\varphi]) \\
& \qquad > \sum_{w' \in S^w_{v(x)}} g^w_{v(x)}(w') Pr^w_{v(x)}(w' \mid [\neg\varphi]).
\end{aligned}
$$

Here is a Levinson model of *in vino veritas.* As before $W = \{w_1, w_2, w_3, w_4\}$, and $v(\textit{me}) = \text{me}$.

$\ulcorner \textit{me}$ wants $b_z \urcorner$, remember, should have a true reading in a situation like this.

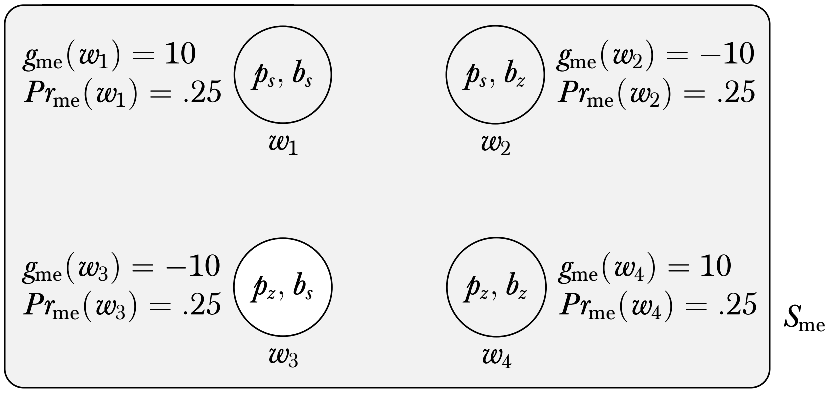
$g_{\text{me}}(w_1) = 10$
$Pr_{\text{me}}(w_1) = .25$
p_s, b_s
w_1
p_s, b_z
w_2
$g_{\text{me}}(w_2) = -10$
$Pr_{\text{me}}(w_2) = .25$
$g_{\text{me}}(w_3) = -10$
$Pr_{\text{me}}(w_3) = .25$
p_z, b_s
w_3
p_z, b_z
w_4
$g_{\text{me}}(w_4) = 10$
$Pr_{\text{me}}(w_4) = .25$
S_{me}

Figure 5.2: A Levinson-style model

What does Levinson's semantics say about it? Well:

$$\begin{aligned}
[\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I} = 1 \quad & \text{iff} \quad EU_{v(me),w_3}([b_z]) > EU_{v(me),w_3}([\neg b_z]) \\
& \text{iff} \quad EU_{\text{me}}([b_z]) > EU_{\text{me}}([\neg b_z]) \\
& \text{iff} \quad \sum_{w' \in S_{\text{me}}} g_{\text{me}}(w') Pr_{\text{me}}(w' \mid [b_z]) \\
& \qquad\qquad > \sum_{w' \in S_{\text{me}}} g_{\text{me}}(w') Pr_{\text{me}}(w' \mid [\neg b_z]) \\
& \text{iff} \quad (10 * 0 + -10 * .5 + -10 * 0 + 10 * .5) \\
& \qquad\qquad > (10 * .5 + -10 * 0 + -10 * .5 + 10 * 0) \\
& \text{iff} \quad 0 > 0.
\end{aligned}$$

Since zero is not greater than zero, Levinson's semantics doesn't help.

Same for (11) and (12); relative to the I defined above, the Levinson truth conditions for (12) runs:

$$\begin{aligned}
[\![p_z \to \textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I} = 1 \quad & \text{iff} \quad \forall w' \in S_{I+p_z}, [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w',I+p_z} = 1 \\
& \text{iff} \quad [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I+p_z} = 1 \\
& \qquad \text{and } [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_4,I+p_z} = 1 \\
& \text{only if} \quad [\![\textit{me} \text{ wants } b_z]\!]^{\mathcal{M},w_3,I+p_z} = 1 \\
& \text{iff} \quad 0 > 0. \quad \text{(Same calculation as above.)}
\end{aligned}$$

Mutatis mutandis for (12). Therefore natural Levinson models of *in vino veritas* do not predict (11) or (12) with respect to information states in which they are naturally asserted.

My proposal

My models are simply Levinson models. The only difference between me and Levinson is the semantic clause for ‘wants’.

Definition. A mixed expected utility function Eu^i_g, relative to a utility function g and sharp information state i, is a function: $\mathcal{P}(W) \to \mathbb{R}$, defined as:

$$EU^i_g(A) := \sum_{w' \in S_i} g(w')Pr_i(w' \mid A \cap S_i)$$

My semantic clause for ‘wants’ is then:

$$\begin{aligned} [\![t \text{ wants } \varphi]\!]^{\mathcal{M},w,I} = 1 \quad & \text{iff} \quad \forall i \in I, EU^i_{g^w_x}([\varphi]) > EU^i_{g^w_x}([\neg\varphi]) \\ & \text{iff} \quad \forall i \in I, \sum_{w' \in S_i} g^w_{v(t)}(w')Pr_i(w' \mid [\varphi]_I) \\ & \qquad > \sum_{w' \in S_i} g^w_{v(t)}(w')Pr_i(w' \mid [\neg\varphi]_I). \end{aligned}$$

Relative to the above information state I, (11) and (12) both come out true. Here’s the derivation of (12). Note that $I + p_s = \{\langle\{w_1, w_2\}, Pr_i^{[p_s]}\rangle\}$, where $Pr_i^{[p_s]}(w_1) = Pr_i^{[p_s]}(w_2) = .5$.

$$\begin{aligned} [\![p_s \to me \text{ wants } b_s]\!]^{\mathcal{M},w_3,I} = 1 \quad & \text{iff} \quad \forall w' \in S_{I+p_s}, [\![me \text{ wants } b_s]\!]^{\mathcal{M},w',I+p_s} = 1 \\ & \text{iff} \quad [\![me \text{ wants } b_s]\!]^{\mathcal{M},w_1,I+p_s} = 1 \\ & \qquad \text{and } [\![me \text{ wants } b_s]\!]^{\mathcal{M},w_2,I+p_s} = 1 \\ & \text{iff} \quad \forall i \in I + p_s, EU^i_{\text{me}}([b_s]) > EU^i_{\text{me}}([\neg b_s]) \\ & \text{iff} \quad 10 * .5 >= -10 * .5 \\ & \text{iff} \quad \top. \end{aligned}$$

Mutatis mutandis for (11).

Relative to any non-trivial information state I' that accepts p_z (i.e. such that $[p_z]_{I'} = S_{I'}$) and the above Levinson model, my semantics predicts $\ulcorner me$ wants $b_z \urcorner$. Any such information state has $S_I = \{w_3, w_4\}$. Thus

$$\begin{aligned} [\![me \text{ wants } b_z]\!]^{\mathcal{M},w_3,I'} = 1 \quad & \text{iff} \quad \forall i \in I', EU^i_{\text{me},w_3}([b_z]) > EU^i_{\text{me},w_3}([\neg b_z]) \\ & \text{iff} \quad 5 > -5 \\ & \text{iff} \quad \top. \end{aligned}$$

The kinds of states of information in which it makes sense to assert the advisory (1)—namely those which accept p_z, as in *in vino veritas*—it's true to ascribe to me a corresponding desire to buy the Zinfandel.

Finally, here is an informational account of consequence:[26]

Definition. $\varphi_1 \ldots \varphi_n \vDash \psi$ iff, for every $\mathcal{M}$, no information state which accepts $\varphi_1 \ldots \varphi_n$ in $\mathcal{M}$ fails to accept ψ in $\mathcal{M}$.

On this definition of consequence, modus ponens comes out valid. Suppose that I accepts φ and $\varphi \to \psi$ relative to some arbitrary $\mathcal{M}$. Since I accepts φ, $I + \varphi = I$. And since I accepts $\varphi \to \psi$, $I + \varphi$ accepts ψ. But then I cannot fail to accept ψ. Thus modus ponens is valid.

However, modus tollens is not valid. This can be shown using the above model and the information state $I = \langle \{w_1, w_2, w_3, w_4\}, Pr_i \rangle$ where $Pr_i(w_j) = .25$ for all j. We've already seen that this information state accepts $p_s \to$ *me* wants b_s. This information state also accepts $\neg($*me* wants $b_s)$, for it assigns b_s the same expected utility as b_z. However, if the actual world is w_3, this model accepts p_s. Thus we have a model and and information state relative to which $\varphi \to \psi$ is accepted, $\neg\psi$ is accepted, but $\neg\varphi$ is not accepted.

[26] See Yalcin (2012a) and Bledin (2014).

Chapter 6

Conclusion

The line pursued in Chapter 5 suggests another possible response to the argument of Chapter 4. Recall that my argument against global case denying involved eschewing radically externalist theories of desire. The idea was that it's largely up to Mal and Ben what they desire when. In particular, Mal, who may be on the opposite side of the world and causally isolated from Ben, can't change what Ben is able to desire merely by changing what she desires. That possibility, I claimed, would be too radically externalist a view to countenance. The content of our desires may supervene widely in the usual ways motivated by standard arguments for content externalism—cases involving natural kinds and singular thoughts, for instance—but Mal's changing her mind about what she desires doesn't seem like the sort of thing that could prevent Ben, no matter how hard he tries, from forming desires that he'd otherwise be able to form. Or so went my argument against global case-denying.

However, as we've seen in the previous chapter, *advisory* desire attributions behave in exactly this way. Recall (1):

(1) He doesn't know it, but he wants the Zinfandel.

(1) can go from true to false depending on the information state at the context of assessment, without changing anything about my underlying psychological state. Thus, it seems, whether I'm in a state of desiring to buy the Zinfandel depends on more than my narrow psychological state; it depends on the information of those assessing (1). If these assessors get information to the effect that my comrades will be happier with the Zinfandel, (1) is true; and

if they get information relative to which they'd prefer the Sauvignon Blanc, it is false. This wide supervenience isn't tied to my ability to *entertain* the contents of the desires, as is plausibly the case with standard cases of content externalism. Instead, my state of desire flips between two different contents, both of which I'm perfectly able to entertain in either case, depending just on what the assessors of the corresponding attribution know about which wine my comrades desire.

How does the relativistic theory of desire attributions developed in the previous chapter treat the case of Mal and Ben? Recall the (putative) case:

> **Mal's strongest desire**: That Ben doesn't get whatever he most strongly desires.
>
> **Ben's strongest desire**: That Mal gets whatever she most strongly desires.

This case is intended to characterize Mal's and Ben's underling states of desire. It is not in the first place about desire *attributions*. But the two are closely connected. Let's explicitly formulate the English-language desire attributions corresponding to the Mal/Ben case:

(2) Mal wants most that Ben doesn't get whatever Ben wants most.

(3) Ben wants most that Mal gets whatever Mal wants most.

Could (2) and (3) be true together, in a single context? The availability of a predictive reading of 'wants' strongly suggests an answer in the affirmative. For if such a reading is available, then the case I originally made for the co-possibility of Mal's and Ben's desires carries over exactly to the attributions of those desires (2) and (3). Mal's and Ben's behavior is best explained in the relevant context by attributing to them desires as in (2) and (3); they explain why Mal is trying to interfere with all of Ben's activities, while Ben is trying to aid Mal in hers. Thus to predict and explain Mal's and Ben's behavior, we do well to attribute (2) and (3). Doing this folk-psychological predictive work is, after all, largely what the predictive use is for.

On an advisory reading, however, it's less clear that (2) and (3) have natural true readings relative to a single context of assessment. Let's start from a simpler case. Suppose Mal is in the state naturally described by (2), and deeply desires the frustration of Ben's desires. Say that you happen to know, as Mal does not, that Ben most strongly desires that the Yankees win their upcoming baseball match. Then you could felicitously utter not only (2), but also something like:

(4) Mal doesn't know it, but she wants the Yankees to lose their upcoming sports match.

This is a typical example of the advisory use at work. You use your information about Ben's love of the Yankees to draw out consequences about what more specific states of affairs would satisfy Mal's more basic desires, and make an advisory attribution on that basis. Since Ben wants the Yankees to win and Mal wants that desire frustrated, she ends up wanting the Yankees to lose, i.e. (4), whether she knows that or not.

Say now that you observe Ben change his desires. Abandoning his ardor for the Yankees, he gets into the state of desire naturally described by (3). Suppose you try to make the same kind of advisory attribution. You would say:

(5) Mal doesn't know it, but she wants it not to be the case that she gets what she most wants.

After all, if (2) and (3) are both the case, a third party should be able to use them to form advisory desire attributions by filling in the conditions that would actually satisfy Mal's more basic desire, whether she knows that or not. Since those conditions involve Ben's desires which in turn involve Mal's desires, (5) results.

Does (5) seem true in the relevant sense here? There is a case to be made that it does. Indeed, it could play a natural intermediate step in an argument directed at Mal that she should change her desires. The following argument sounds pretty convincing:

> Look, you shouldn't most want Ben not to get what he most wants. You may not have known this, but Ben actually most wants that you *get* what you most want! So in a sense, you most want yourself not to get what you most want, and that's crazy.

This monologue uses (5) in the course of arguing that Mal should desire something different from what she actually desires. This argument makes sense only if we grant that, originally, Mal *does* desire the thing the argument is trying to get her not to desire, namely, the frustration of all of Ben's desires.

If that's right, then there is a case to be made that, even on an advisory reading, (2) and (3) can be true.

However, one might worry that this is not taking the advisory reading far enough. After all, the data of Chapter 5 suggests that *anytime* a third party has information according to which the desirer's more basic aims are frustrated by φ, she does not really desire φ, no matter how much it may seem like she does. This seems to be the case here: whatever Mal's basic aims might plausibly be said to be, it seems like forming a strongest desire to the effect that Ben not get what he most strongly desires cannot be conducive to Mal's actual well-being (given what Ben desires). If that's right, then on the advisory reading, Mal never desired that all along. As soon as Ben started desiring the satisfaction of Mal's strongest desire, Mal would be badly off desiring the frustration of Ben's. And since she'd be badly off doing that, she doesn't really want it in the advisory sense.

There's definite plausibility to the idea that Mal and Ben each must have *some* more basic aims than attributed in (2) and (3), aims which are not actually furthered in the relevant case by the formal satisfaction of the downstream paradoxical desires that they each seem to have. However, this runs counter to the spirit of the case originally described in Chapter 4. In that description, Mal was supposed to fundamentally and primitively most strongly desire the frustration of Ben's strongest desire—not in virtue of beliefs about what other of her aims that might further, but just as a basic fact about her fundamental aims. Her well being is *characterized by*, and depends exclusively on, the frustration of Ben's desires. Thus she has no deeper aims in virtue of which she's formed this desire; this desire exhausts her deepest aims.

This is, of course, highly implausible as a description of any actual agent. Perhaps such an agent is incompatible with certain principles of psychology. The question, though, is whether *logic alone* can exclude the possibility of such agent. And however we conceive of logic, it would be very surprising to learn that logic alone forbid something substantive about what most basic desires agents can have. This is especially so if one is a global case denier—for then what logic forbids me from desiring depends on what others far away from me happen to desire.

Thus, the prospects for using the data of Chapter 5 to undermine the dialectic of Chapter 4 seem, if not hopeless, at least dim. One would have to both deny the possibility of a predictive reading of "wants"—which, as we saw in 5.6, is a live but not settled option—and make the case that *no* *possible* pair of agents is such that, relative to an advisory reading, something like (2) and (3) are both true. This is quite a narrow line to walk. Thus, a non-classical approach to the paradoxical situation of Chapter 4 remains a more robust and well-motivated solution than an appeal to the non-existence of a predictive "wants" and the logical impossibility of having basic desires like those of Mal and Ben.

This is by no means the end of the story. A non-classical model of desire attributions that allows for the simultaneous truth of (2) and (3) remains forthcoming, and will have to await future work. I hope to have made the case here that such a thing is worth creating.

Bibliography

Anderson, C. Anthony (1983). “The Paradox of the Knower”. In: *Journal of Philosophy* 80.6, pp. 338–355.

Bacon, Andrew (forthcoming). “Radical Anti-Disquotationalism”. In: *Philosophical Perspectives.*

— (2013). “Non-Classical Metatheory for Non-Classical Logics”. In: *Journal of Philosophical Logic* 42.2, pp. 335–355.

Bacon, Andrew, John Hawthorne, and Gabriel Uzquiano (2016). “Higher-Order Free Logic and the Prior-Kaplan Paradox”. In: *Canadian Journal of Philosophy* 46.4-5, pp. 493–541.

Bacon, Andrew and Gabriel Uzquiano (2018). “Some Results on the Limits of Thought”. In: *Journal of Philosophical Logic*, pp. 1–9.

Bledin, Justin (2014). “Logic Informed”. In: *Mind* 123.490, pp. 277–316.

Boolos, George S., John P. Burgess, and Richard C. Jeffrey (2007). *Computability and Logic.* 5th. Cambridge: Cambridge University Press.

Bradley, Richard and H. Orri Stefánsson (2016). “Desire, Expectation and Invariance”. In: *Mind* 125.499, pp. 691–725.

Buchak, Lara (2013). *Risk and Rationality.* OUP Oxford.

Burge, Tyler (1979a). “Individualism and the Mental”. In: *Midwest Studies in Philosophy IV: Studies in Metaphysics.* Ed. by Peter French, Theodore E. Uehling, Jr., and Howard K. Wettstein. Minneapolis, MN: University of Minnesota Press, pp. 73–121.

— (1979b). “Semantical Paradox”. In: *Journal of Philosophy* 76.4, pp. 169–198.

Büring, Daniel (2003). “To want is to want to be there: A note on Levinson 2003”. URL: http://linguistics.ucla.edu/general/Conf/LaBretesche/papers/buring.pdf.

Caie, Michael (2012). “Belief and Indeterminacy”. In: *Philosophical Review* 121.1, pp. 1–54.

Callard, Agnes (2017). “Everyone Desires the Good: Socrates’ Protreptic Theory of Desire”. In: *Review of Metaphysics* 70.4.

Chalmers, David (1995). “Minds, Machines, And Mathematics A Review of Shadows of the Mind by Roger Penrose”. In: *Psyche* 2.

Condoravdi, Cleo and Sven Lauer (Nov. 2016). “Anankastic conditionals are just conditionals”. In: *Semantics and Pragmatics* 9.8, pp. 1–69. DOI: 10.3765/sp.9.8.

Cross, Charles B. (2001). “The Paradox of the Knower Without Epistemic Closure”. In: *Mind* 110.438, pp. 319–333.

— (2004). “More on the Paradox of the Knower Without Epistemic Closure”. In: *Mind* 113.449, pp. 109–114.

Davidson, Donald (1975). “Thought and Talk”. In: *Mind and Language*. Ed. by Samuel D. Guttenplan. Clarendon Press, pp. 1975–7.

Davis, Wayne A. (1984). “The Two Senses of Desire”. In: *Philosophical Studies* 45.2, pp. 181–195.

Dowell, Janice J. L. (2011). “A Flexible Contextualist Account of Epistemic Modals”. In: *Philosophers’ Imprint* 11.14, pp. 1–25.

— (2013). “Flexible Contextualism About Deontic Modals: A Puzzle About Information-Sensitivity”. In: *Inquiry* 56.2-3, pp. 149–178.

Dummett, Michael (1974). “Frege: Philosophy of Language”. In: *Philosophical Quarterly* 24.97, pp. 349–359.

Égré, Paul (2005). “The Knower Paradox in the Light of Provability Interpretations of Modal Logic”. In: *Journal of Logic, Language and Information* 14.1, pp. 13–48.

Field, Hartry (1978). “Mental Representation”. In: *Erkenntnis* 13.July, pp. 9–61.

— (2003). “The Semantic Paradoxes and the Paradoxes of Vagueness”. In: *Liars and Heaps: New Essays on Paradox*. Ed. by J. C. Beall. Oxford University Press, pp. 262–311.

— (2008). *Saving Truth from Paradox*. Oxford University Press.

— (2016). “Indicative Conditionals, Restricted Quantifiers and Naive Truth”. In: *Review of Symbolic Logic*, pp. 1–28.

von Fintel, Kai (1999). “NPI Licensing, Strawson Entailment, and Context Dependency”. In: *Journal of Semantics* 16.2, pp. 97–148.

— (2012). “The best we can (expect to) get? Challenges to the classic semantics for deontic modals”. In: *Central APA*.

Fodor, Jerry A. (1975). *The Language of Thought*. Harvard University Press.

Geach, Peter (1957). *Mental Acts*. Routledge and Kegan Paul.

Glanzberg, Michael (2001). “The Liar in Context”. In: *Philosophical Studies* 3, pp. 217–251.

— (2004). “A Contextual-Hierarchical Approach to Truth and the Liar Paradox”. In: *Journal of Philosophical Logic* 33.1, pp. 27–88.

Goble, Lou (1996). “Utilitarian Deontic Logic”. In: *Philosophical Studies* 82.3, pp. 317–357.

Gödel, Kurt (1995). “Some basic theorems on the foundations of mathematics and their implications”. In: *Collected Works, Vol. III: Unpublished Essays and Lectures.* Ed. by Solomon Feferman. Oxford University Press, pp. 304–323.

Halbach, V. and P. Welch (2009). “Necessities and Necessary Truths: A Prolegomenon to the Use of Modal Logic in the Analysis of Intensional Notions”. In: *Mind* 118.469, pp. 71–100.

Halbach, Volker and Albert Visser (2014a). “Self-Reference in Arithmetic I”. In: *Review of Symbolic Logic* 7.4, pp. 671–691.

— (2014b). “Self-Reference in Arithmetic II”. In: *Review of Symbolic Logic* 7.4, pp. 692–712.

Heim, Irene (1992). “Presupposition Projection and the Semantics of Attitude Verbs”. In: *Journal of Semantics* 9.3, pp. 183–221.

Holliday, Wesley H. and Thomas F. Icard (2013). “Measure Semantics and Qualitative Semantics for Epistemic Modals”. In: *Proceedings of SALT 23*, pp. 514–534.

Holliday, Wesley H., Thomas F. Icard, and Matthew Harrison-Trainor (2017). “Preferential Structures for Comparative Probabilistic Reasoning”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.*

Jerzak, Ethan (2019). “NonClassical Knowledge”. In: *Philosophy and Phenomenological Research* 98.1, pp. 190–220. DOI: 10.1111/phpr.12448.

Kaplan, David and Richard Montague (1960). “A Paradox Regained”. In: *Notre Dame Journal of Formal Logic* 1, pp. 79–90.

Koellner, Peter (2016). “On Gödel’s Disjunction”. In: *Godel’s Disjunction: The Scope and Limits of Mathematical Knowledge.* Ed. by Leon Horsten and Philip Welch. Oxford University Press Uk.

Kolodny, Niko and John MacFarlane (2010). “Ifs and Oughts”. In: *The Journal of Philosophy* 107.3.

Kripke, Saul (1975). “Outline of a Theory of Truth”. In: *Journal of Philosophy* 72, pp. 690–716.

Kyburg Jr, Henry E. (1961). *Probability and the Logic of Rational Belief.* Vol. 34. Wesleyan University Press, pp. 283–285.

Lassiter, Daniel (2011). "Measurement and Modality: the Scaler Basis of Modal Semantics". PhD thesis. New York University.

Levinson, Dmitry (2003). "Probabilistic model-theoretic semantics for *want*". In: *SALT*. Vol. 13, pp. 222–239.

Lewis, David (1979). "Attitudes *De Dicto* and *De Se*". In: *Philosophical Review* 88, pp. 513–43.

— (1988). "Desire as Belief". In: *Mind* 97.418, pp. 323–32.

MacFarlane, John (2014). *Assessment Sensitivity: Relative Truth and its Applications*. Oxford University Press.

Magnus, P. D. et al. (2018). *Forall X: Calgary Remix*. CreateSpace.

Maitzen, Stephen (1998). "The Knower Paradox and Epistemic Closure". In: *Synthese* 114.2, pp. 337–354.

Makinson, D. C. (1965). "The Paradox of the Preface". In: *Analysis* 25, pp. 205–207.

Maudlin, Tim (2004). *Truth and Paradox*. Oxford University Press.

Penrose, Roger (1994). *Shadows of the Mind*. Oxford University Press.

Phillips-Brown, Milo (2017). "I want to, but…" In: *Proceedings of Sinn und Bedeutung* 21.

Picollo, Lavinia and Thomas Schindler (July 2017). "Disquotation and Infinite Conjunctions". In: *Erkenntnis*. ISSN: 1572-8420. DOI: 10.1007/s10670-017-9919-x. URL: https://doi.org/10.1007/s10670-017-9919-x.

Priest, Graham (1987). *In Contradiction: A Study of the Transconsistent*. Dordrecht: Martinus Nijhoff.

Prior, A. N. (1961). "On a Family of Paradoxes". In: *Notre Dame Journal of Formal Logic* 2.1, pp. 16–32.

Putnam, Hilary (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1951). "Two Dogmas of Empiricism". In: *Philosophical Review* 60.1, pp. 20–43.

Ramsey, F. P. (1927). "Facts and Propositions". In: *Proceedings of the Aristotelian Society* 7.1, pp. 153–170.

van Rooij, Robert (1999). "Some Analyses of Pro-Attitudes". In: *Logic, Game Theory, and Social Choice*. Ed. by H. de Swart. Tilburg University Press, pp. 263–279.

Rooryck, Johan (2017). "Between desire and necessity: the complementarity of 'want' and 'need'". In: *Crossroads Semantics: Computation, Experiment, and Grammar*. Ed. by Hilke Reckman et al. John Benjamins Publishing Company, pp. 263–279.

Russell, Gillian (2018). "Logical Nihilism: Could There Be No Logic?" In: *Philosophical Issues* 28.1, pp. 308–324. doi: 10.1111/phis.12127.

Sainsbury, R. M. (1995). *Paradoxes.* Cambridge University Press.

Schlenker, Philippe (2004). "Context of Thought and Context of Utterance: A Note on Free Indirect Discourse and the Historical Present". In: *Mind and Language* 19.3, pp. 279–304.

Schroeder, Mark (2011). "Ought, Agents, and Actions". In: *Philosophical Review* 120.1, pp. 1–41.

Sellars, Wilfrid S. and Roderick M. Chisholm (1957). "Intentionality and the Mental: A Correspondence". In: *Minnesota Studies in the Philosophy of Science* 2, pp. 507–39.

Shapiro, Stewart (2003). "Mechanism, Truth, and Penrose's New Argument". In: *Journal of Philosophical Logic* 32.1, pp. 19–42.

Stalnaker, Robert (1984). *Inquiry.* Cambridge, MA: MIT Press.

Stern, Johannes (2014). "Montague's Theorem and Modal Logic". In: *Erkenntnis* 79.3, pp. 551–570.

Tarski, Alfred (1936). "Der Wahrheitsbegriff in den formalisierten Sprachen". In: *Studia Philosophica* 1, pp. 261–405.

Thomason, Richmond H. (1981). "Deontic Logic as Founded on Tense Logic". English. In: *New Studies in Deontic Logic.* Ed. by Risto Hilpinen. Vol. 152. Synthese Library. Springer Netherlands, pp. 165–176. isbn: 978-90-277-1346-9. doi: 10.1007/978-94-009-8484-4_7. url: http://dx.doi.org/10.1007/978-94-009-8484-4_7.

Uzquiano, Gabriel (2004). "The Paradox of the Knower Without Epistemic Closure?" In: *Mind* 113.449, pp. 95–107.

Villalta, Elisabeth (2000). "Spanish subjunctive clauses require ordered alternatives". In: *SALT.* Vol. 10.

Wang, Hao (1997). *A Logical Journey: From Gödel to Philosophy.* A Bradford Book.

Whittle, Bruno (2017). "Self-Referential Propositions". In: *Synthese* 194.12, pp. 5023–5037.

Yalcin, Seth (2007). "Epistemic Modals". In: *Mind* 116.464, pp. 983–1026.

— (2012a). "A Counterexample to Modus Tollens". In: *Journal of Philosophical Logic* 41.6, pp. 1001–1024.

— (2012b). "Bayesian Expressivism". In: *Proceedings of the Aristotelian Society* 112.2pt2, pp. 123–160.

Yalcin, Seth (2012c). “Context Probabilism”. In: *Logic, Language and Meaning.* Ed. by M. Aloni. Vol. 7218. Lecture Notes in Computer Science. Springer, pp. 12–21.