Non-Classical Knowledge

Forthcoming in *Philosophy and Phenomenological Research*

Ethan Jerzak

May 27, 2017

Abstract

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, that we can know that knowledge is factive, and that we can use logical/mathematical reasoning to extend our knowledge via very weak single-premise closure principles. I argue that classical logic, not any of these epistemic principles, is the culprit. I develop a consistent theory validating all these principles by combining Hartry Field's theory of truth with a modal enrichment developed for a different purpose by Michael Caie. The only casualty is classical logic: the theory avoids paradox by using a weaker-than-classical K_3 logic.

I then assess the philosophical merits of this approach. I argue that, unlike the traditional semantic paradoxes involving extensional notions like truth, its plausibility depends on the *way* in which sentences are referred to—whether in natural languages via direct sentential reference, or in mathematical theories via indirect sentential reference by Gödel coding. In particular, I argue that from the perspective of natural language, my non-classical treatment of knowledge as a predicate is plausible, while from the perspective of mathematical theories, its plausibility depends on unresolved questions about the limits of our idealized deductive capacities.

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 (§1). 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 (§2). 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 (§3). 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 (§4). 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.

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

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

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]

$$T(\ulcorner\varphi\urcorner) \supset \varphi; \tag{T-out}$$

[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.

$$\varphi \supset T(\ulcorner \varphi \urcorner). \tag{T-in}$$

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, 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.

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. One of the things we talk about in English

[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 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

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 $K(\ulcorner\varphi\urcorner)$ hold just in case some rational agent α knows the sentence φ.[4]

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

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

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

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

i	$K(\ulcorner\varphi\urcorner)$		
j	φ		
...	...	*(This subproof can use nothing derived from premises;*	
k	ψ	*only the rules and axioms of theory X.)*	(SPC)
$k+1$	$K(\ulcorner\psi\urcorner)$	SPC_X, i, j-k	

(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 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:

[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.

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

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]

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 $\{(FACT_g), (KFACT_g)\}$, 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]

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

[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.

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

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 (§2.1), and a novel conditional called the Field conditional (§2.2). In §2.3, 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.

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:

$$\begin{aligned}
[\![\phi \wedge \psi]\!] &= \min\{\ [\![\phi]\!],\ [\![\psi]\!]\} \\
[\![\phi \vee \psi]\!] &= \max\{\ [\![\phi]\!],\ [\![\psi]\!]\} \\
[\![\neg\phi]\!] &= 1 -\ [\![\phi]\!] \\
\varphi \supset \psi &:= \neg\varphi \vee \psi
\end{aligned}$$

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 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.

[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.

[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.

The classical rules that are not valid in K_3 are those which allow us to exit Fitchean subderivations: 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 $\bot$, 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 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 of this paper 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

[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.

[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.

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 **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 $\frac{1}{2}$.

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.

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

[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.

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 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 $\mathrm{K_3FT}$ ('$\mathrm{K_3}$' for $\mathrm{K_3}$, '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 2.1. *All instances of the following schema are valid:*

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

Proof. See Appendix A. □

Proposition 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. □

[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.

Proposition 2.3. *The following rule is valid:*

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

φ	
...	
ψ	*(Only axioms and rules of K_3FT.)*

$\Box T(\ulcorner\psi\urcorner)$ *Closure*$_{K_3FT}$.

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 $\Box$ 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	$K(\ulcorner g \urcorner)$	
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 $\perp$ in K_3FT 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_3FT is sound on it. Since $\perp$ 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 $\perp$. 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 $\perp$ from (FACT), (KFACT), and (SPC_{K_3FT}) in K_3FT, 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.

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 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 (§3.1) 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 (§3.2) 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 in §4 to natural languages, where the complications of indirect sentential reference via Gödel coding can be avoided.

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

[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.

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 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; Quine corners 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 $\ulcorner g \urcorner$

[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.

is some numeral (which numeral it is depends on the coding). So in the actual object language, (FACT) will really look something like:

$$K(\overline{587}) \rightarrow \exists x(x = 0) \tag{1}$$

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 ⟨some formula in $\mathcal{L}_{PA+K}$⟩. 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:

$$K(\overline{3832}) \tag{2}$$

where $\overline{3832}$ decodes (relative to the coding scheme we chose) to (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 Quine corners. (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,

$$PA \vdash K(\overline{587}) \rightarrow \exists x(x = 0). \tag{3}$$

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

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

$$K(K\varphi \rightarrow \varphi) \tag{4}$$

does not suffer from the same problem. Here φ is being used twice, not used once and mentioned once (as in (KFACT)). So (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.

'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.

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), 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) 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 $\mathbf{F}$ be an arbitrary finitely axiomatizable theory containing PA in $\mathcal{L}_{PA+K}$, let $\mathbf{K}$ be the set of sentences in this language ideally knowable by us, and let $\mathbf{T}$ be the set of true sentences in this language. We restrict to theories $\mathbf{F}$ which are themselves absolutely knowable and assume that what's absolutely knowable is true, which

automatically yields the following relationship between these sets:

$$\mathbf{F} \subseteq \mathbf{K} \subseteq \mathbf{T} \tag{5}$$

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 in §4 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 in

§2.3, 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,

$$\mathbf{F} \nvdash \neg \mathrm{Prov}_{\mathbf{F}}(\ulcorner 0{=}1 \urcorner) \tag{6}$$

however, as an instance of (KFACT),

$$K(\ulcorner K(\ulcorner 0 = 1 \urcorner) \supset 0 = 1 \urcorner) \tag{7}$$

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

$$\mathbf{F} \vdash K(\ulcorner 0 = 1 \urcorner) \supset 0 = 1. \tag{8}$$

But we can't have

$$\mathbf{F} \vdash \mathrm{Prov}_{\mathbf{F}}(\ulcorner 0 = 1 \urcorner) \supset 0 = 1, \tag{9}$$

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

because that would contradict (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* $\mathbf{F}$ aligns with what I know: We'd get an actual inconsistency by adding $K(\ulcorner \text{Prov}_{\mathbf{F}}(\ulcorner \varphi \urcorner) \supset K(\ulcorner \varphi \urcorner) \urcorner)$, which combined with (8) would entail (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.

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. 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 in §3, 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:

[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.

(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 $\bot$ in §1.2. (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:

$$K(T(\ulcorner\varphi\urcorner)) \leftrightarrow K(\varphi) \tag{10}$$

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:

$$K(\ulcorner T(\imath x S(x))\urcorner) \wedge \neg K(\imath x S(x)) \tag{11}$$

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:

$$K(\ulcorner T(*)\urcorner) \wedge \neg K(*) \tag{12}$$

Here, though, (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. (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 (11) could be true doesn't hold with (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, since (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, (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).

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 aban-

doning (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, 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 doesn't 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

[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 without semantic/epistemic vocabulary go through on my theory.

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 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.[21]

References

Anderson, C. Anthony (1983). "The Paradox of the Knower". In: *Journal of Philosophy* 80.6, pp. 338–355.

Bacon, Andrew (2013). "Non-Classical Metatheory for Non-Classical Logics". In: *Journal of Philosophical Logic* 42.2, pp. 335–355.

Caie, Michael (2012). "Belief and Indeterminacy". In: *Philosophical Review* 121.1, pp. 1–54.

Chalmers, David (1995). "Minds, Machines, And Mathematics A Review of Shadows of the Mind by Roger Penrose". In: *Psyche* 2.

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.

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.

[21]Thanks to Chloé de Canson, Wes Holliday, Hannes Leitgeb, John MacFarlane, and Seth Yalcin for comments on drafts. Thanks also to audiences at Berkeley, ESSLLI, and the MCMP.

Field, Hartry (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.

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.

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.

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.

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.

Priest, Graham (1987). *In Contradiction: A Study of the Transconsistent*. Dordrecht: Martinus Nijhoff.

Sainsbury, R. M. (1995). *Paradoxes*. Cambridge University Press.

Shapiro, Stewart (2003). "Mechanism, Truth, and Penrose's New Argument". In: *Journal of Philosophical Logic* 32.1, pp. 19–42.

Stern, Johannes (2014). "Montague's Theorem and Modal Logic". In: *Erkenntnis* 79.3, pp. 551–570.

Uzquiano, Gabriel (2004). "The Paradox of the Knower Without Epistemic Closure?" In: *Mind* 113.449, pp. 95–107.

Wang, Hao (1997). *A Logical Journey: From Gödel to Philosophy*. A Bradford Book.

Appendix A: The Theory

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!):

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

$[\![P(t_i)]\!]^{\mathcal{M},w,\mu} = 1$ iff $[t_i]_{\mathcal{M},\mu} \in v(w, P)$;
$[\![P(t_i)]\!]^{\mathcal{M},w,\mu} = 0$ otherwise;

$[\![\phi \wedge \psi]\!]^{\mathcal{M},w,\mu} = \min\{\ [\![\phi]\!]^{\mathcal{M},w,\mu},\ [\![\psi]\!]^{\mathcal{M},w,\mu}\};$
$[\![\phi \vee \psi]\!]^{\mathcal{M},w,\mu} = \max\{\ [\![\phi]\!]^{\mathcal{M},w,\mu},\ [\![\psi]\!]^{\mathcal{M},w,\mu}\};$
$[\![\neg\phi]\!]^{\mathcal{M},w,\mu} = 1 - \ [\![\phi]\!]^{\mathcal{M},w,\mu};$
$[\![\Box\varphi]\!]^{\mathcal{M},w,\mu} = \min\{\ [\![\varphi]\!]^{\mathcal{M},v,\mu} : wRv\};$
$[\![\forall x_i\varphi]\!]^{\mathcal{M},w,\mu} = \min\{\ [\![\varphi]\!]^{\mathcal{M},w,\mu'} : \mu' =_{x_i} \mu\ \}$.

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 $\rightarrow$ 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 $\rightarrow$.

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 $\rightarrow$ 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:

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

Let $T^{w+}_0 = \emptyset$, and let $T^{w+}_{\alpha+1}$ be the set of sentences φ such that $[\![\varphi]\!]^{\mathcal{M}^{+\alpha},w} = 1$. Similarly let T^{w-}_0 be the empty set, and let $T^{w-}_{\alpha+1}$ be the set of sentences φ such that $[\![\varphi]\!]^{\mathcal{M}^{+\alpha},w} = 0$, as well as all non-sentences. At a limit ordinal λ, T^{w+}_λ is the set of sentences φ such that, for some $\beta < \lambda$, $[\![T(t_i)]\!]^{\mathcal{M}^{+\beta},w} = 1$ (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 $[\![\varphi]\!]^{\mathcal{M}^{+\sigma}, w} = 1$ 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.[22]

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 is pretty much what you would expect. For the class $\mathbb{M}$ 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 validity, the Fundamental Theorem guarantees that the corresponding logic remains K_3 and not something weird: all individual Field-plus-Kripke models have a K_3 logic, and the formulas are all together stabilized at the acceptable Field-plus-Kripke 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 $\rightarrow$.

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

[22] 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 $\frac{1}{2}$.

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 $\frac{1}{2}$.[23]

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 cannot be strictly greater than that of $T(\ulcorner \varphi \urcorner)$ in w. But at any world in an acceptable Field-plus-Kripke 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 $\Box T(\ulcorner \varphi \urcorner) \rightarrow \varphi$ stabilizes at 1 for all worlds and models. Thus $\models_{\mathbb{M}} K(\ulcorner \varphi \urcorner) \rightarrow \varphi$. □

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

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

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:*

$\Box T(\ulcorner \varphi \urcorner)$	
φ	
⋮	
ψ	(Only axioms and rules of K_3FT.)
$\Box T(\ulcorner \psi \urcorner)$	Closure$_{K_3FT}$.

[23] 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 $\frac{1}{2}$.

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.[24] □

Appendix B: Direct self-reference

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.[25] 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 Quine corners 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

[24] 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 $K(\ulcorner\varphi_1\urcorner), \ldots, K(\ulcorner\varphi_n\urcorner) \vdash_{K_3FT} K(\ulcorner\psi\urcorner)$.

[25] "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.

we want $\ulcorner\cdot\urcorner$ always to be well-defined, we saddle our models with the following restriction: Fix a particular enumeration of the 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. 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

$$\begin{aligned}
[\![\neg T(n_i)]\!]^{\mathcal{M},w} &= 1 - [\![T(n_i)]\!]^{\mathcal{M},w} && \text{Semantics of } \neg \\
&= 1 - [\![[n_i]_{\mathcal{M}}]\!]^{\mathcal{M},w} && \text{By T-equiv. (Only makes sense when } [n_i] \text{ is a sentence)} \\
&= 1 - [\![\neg T(n_i)]\!]^{\mathcal{M},w} && \text{Because of what } n_i \text{ denotes.}
\end{aligned}$$

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 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:

[26] This difference in our interpretation of the Quine corners will change what the actual proofs in the object language look like, once the Quine corners 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 Quine corners look exactly the same; but when you interpret the Quine corners and look at the honest-to-God object-language forms of the proof, they look a bit different.

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

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:

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

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.