3-24-2021

Epistemic Blindspots and Godel's First Incompleteness Theorem

Phillip Allen Lane

Epistemic Blindspots and Gödel's First Incompleteness Theorem

by

Phillip Allen Lane

An Honors Capstone

submitted in partial fulfillment of the requirements

for the Honors Diploma

to

The Honors College

of

The University of Alabama in Huntsville

March 24, 2021

Honors Capstone Directors: Dr. Andrew Cling and Dr. Nicholaos Jones

Professors of Philosophy

Student	4/20/2021 Date
Director (Dr. Cling)	Digitally signed by Andrew D Cling DN: cn=Andrew D Cling, o=UAH, ou=AHS Deans Office, email=clinga@uah.edu, c=US Date: 2021.04.05 10:51:55 -05'00' Date
Director (Dr. Jones)	Digitally signed by Nicholaos Jones Date: 2021.04.20 16:23:26 -05'00' Date
Department Chair	Digitally signed by Nicholaos Jones Date: 2021.04.20 16:23:42 -05'00' Date
Honors College Dean	William S Wilkerson Digitally signed by William S Wilkerson Date: 2021.04.23 08:48:32 -05'00' Date

HONORS COLLEGE
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

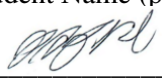
Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

Honors Thesis Copyright Permission

This form must be signed by the student and submitted as a bound part of the thesis.

In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

Phillip A. Lane
_____
Student Name (printed)


_____
Student Signature

4/20/21
_____
Date

Table of Contents

Acknowledgement:

I would like to give special thanks first and foremost to Andrew Cling and Nicholaos Jones for helping me through the entire process, from conceptualization to completion, of writing a philosophical essay. I would also like to thank my parents, Preston and Lisa Lane, for inspiring me to be the best I can possibly be, and Mark Reuter, for helping verify some proofs demonstrated in this essay.

Abstract

In this essay, I endeavor to demonstrate that Gödel's first incompleteness theorem has an implication about epistemic blindspots. "Blindspot" is a term used by Roy Sorensen, and is defined as a true but unknowable proposition. It is inspired by the notion of a visual blind spot, which is an area of our vision that we are blind to due to the optic nerve, and is filled in by the brain using surrounding information.

This essay aims to introduce the notion of "Gödel blindspots," which are propositions that are true yet unknown due to the diagonal lemma. Gödel blindspots only occur in theories of arithmetic in which "knowledge" is a predicate, which is necessary to properly utilize the diagonal lemma. Nonetheless, blindspots of a similar nature, that being self-referential sentences that deny a predicate of themselves, exist in natural language, such as in sentences such as "Necessarily, I am not known." These types of sentences provide a strong case against the K Principle, the primary premise used in Fitch's paradox of knowability to derive the conclusion of epistemic trivialism. Implications of Fitch's paradox are discussed in-depth.

Introduction

Epistemic blindspots are true but unknowable propositions. They are inspired by the notion of a visual blind spot, which is an area of our vision that we cannot see and is "filled in" using surrounding visual information. While we are not usually aware of our visual blind spots, it is possible to witness an item in the visual field disappear if it falls into the blind spot. Likewise, while we are not often aware of epistemic blindspots, we may notice individual propositions that that are unknowable if we know where to look. Sorensen, in his book *Blindspots*, gives an exercise to witness one's visual blind spot. I have recreated it here. Hold the page about 18 inches away from your eyes. Close your left eye, then look at the left dot. You will see the right dot disappear on the paper.

Much like visual blind spots, where you can find the other dot by shifting your gaze or looking from a different position, many blindspots can be "spotted" by being analyzed at a different time, by being analyzed by a different person, or by using different means. There are many types of blindspots, such as blindspots about belief (such as the sentence-form, "*P* but I don't believe *P*," which cannot be rationally believed by the one who utters it), blindspots about knowledge (such as the sentence-form "*P* but I don't know *P*," which cannot be known by the one who utters it), and absolute variants of each of these (such as "*P* but no one knows that *P*," which cannot be known by anybody).

Gödel's first incompleteness theorem is a theorem in metamathematics that asserts the existence of consistent but unprovable propositions in the language of arithmetic. It is done by virtue of a diagonal argument. First, one introduces a Gödel numbering, which is an injection from the sentences of arithmetic to numerals, or concatenations of numerals. One then proves that a

function exists that computes the diagonalization of a sentence of arithmetic. You can then use that to prove the existence of certain self-referential sentences. This proves that arithmetic isn't decidable. Then, by proving that all axiomatizable and complete theories are decidable, you prove that arithmetic is incomplete (i.e. that there exist consistent yet unprovable propositions). Arguably, the most crucial step in this proof is proving the existence of certain self-referential sentences. This section of the proof is deemed the "diagonal lemma," or less commonly, the "fixed-point lemma," which will be what we will mostly concern ourselves with.

This essay aims to expand the notion of "blindspots" into the realm of formal epistemology. Using Gödel's first incompleteness theorem, and referring to "knowledge" as a formula represented in arithmetic, I will prove that a different kind of blindspot exists—a blindspot that doesn't necessarily generate "unknowable" propositions, but blindspots in the sense that there are necessarily unknown propositions. I say "blindspot" in the $\Box \exists$ sense, rather than the $\exists \Box$ sense (Humberstone 2019), such that these blindspots necessarily exist, not that there exist sentences which are necessarily blindspots. I will first augment Gödel's first incompleteness theorem, by following it up until the diagonal lemma, then instantiating the quantified fixed-point sentence with the negation of a factive knowledge predicate. Next, I will examine Fitch's paradox of knowability and analyze the K Principle (the central premise of the paradox) through the lens of knowledge represented in arithmetic. Then, I will identify some self-referential blindspots in natural language, as inspired by the central theorem of the paper. Finally, I will revisit Fitch's paradox one final time after examining self-referential blindspots that assert that they are not known.

The Proof Expressed

The primary proof hinges on proving Gödel's first incompleteness theorem up to the diagonal lemma, applying the diagonal lemma to demonstrate that there exists a sentence ϕ such that $\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)$ where $Know$ can have several properties, but must include factivity: $Know(\ulcorner \phi \urcorner) \rightarrow \phi$. Finally, through a simple derivation in first-order logic, I show that there necessarily exists a ϕ such that $\phi \,\&\, \neg Know(\ulcorner \phi \urcorner)$. The proof up through the diagonal lemma will be adapted from Boolos and Jeffrey. The intention behind restating the proof here is so that the reader is familiar with all of the premises and context behind the final result, as to avoid misapplication of this result in formal epistemology.

Preliminary Material

The most important thing to understand before proving the diagonal lemma is to have an understanding of Gödel numberings. Gödel numberings are a way of assigning a unique number to each sentence of arithmetic. In other words, we create an injection of the sentences of arithmetic onto the natural numbers (each natural number included in the mapping is associated with at most one sentence of arithmetic). The mapping is from the following table of symbols:

(	)	&	∃	x_0	f_0^0	f_0^1	f_0^2	...	A_0^0	A_0^1	A_0^2	...
	,	∨	∀	x_1	f_1^0	f_1^1	f_1^2	...	A_1^0	A_1^1	A_1^2	...
		¬		x_2	f_2^0	f_2^1	f_2^2	...	A_2^0	A_2^1	A_2^2	...
		↔		.	.	.	.		.	.	.	
		→		.	.	.	.		.	.	.	
				.	.	.	.		.	.	.	

To the corresponding numbers in the following table:

1	2	3	4	5	6	68	688	...	7	78	788	...
	29	39	49	59	69	689	6889	...	79	789	7889	...
		399		599	699	6899	68899	...	799	7899	78899	...
		3999		.	.	.	.		.	.	.	
		39999		.	.	.	.		.	.	.	
				.	.	.	.		.	.	.	

As you can see, each symbol is assigned a unique natural number. We can look at a number on the bottom table and identify which symbol it represents. For instance, the number 399 can be observed to represent the symbol "¬". Next, we shall assign a couple of symbols of logic to symbols of arithmetic. Boolos and Jeffrey call these "'conventions' about the identity of certain symbols." The assignments are as follows:

$$x_0 = x$$

$$x_1 = y$$

$$f_0^0 = 0$$

$$f_0^1 = succ$$

$$f_0^2 = +$$

$$f_1^2 = \cdot$$

$$A_0^2 = =$$

Therefore, the Gödel number of $succ$ is 68, and the Gödel number of $\cdot$ is 6889.

Recall that we said earlier that we were creating an injection of all *sentences* of arithmetic to the natural numbers. So far, we have only created an injection of the *symbols* of arithmetic to the naturals. So, let's extend our system to include all sentences, and not just individual symbols. To do that, we rely on the strategy by which we've defined our mapping so far. The way in which we defined the mapping allows us to simply define the Gödel number of a sentence of arithmetic

as the concatenation of all of the symbols' Gödel numbers. An example: $s0 + s0 = ss0$ first becomes $A_0^2 f_0^2 f_0^1 f_0^0 f_0^1 f_0^0 f_0^1 f_0^1 f_0^0$, which then becomes 788688668866886886.

Finally, we will discuss the diagonalization of a sentence. A quick note about syntax: if the Gödel number of an expression ϕ is n, then we will say that $\ulcorner \phi \urcorner = \mathbf{n}$ where $\mathbf{n}$ is a series of symbols in the language. With that said, the diagonalization of an expression ϕ is the sentence:

$$(\exists x)\big(x = \ulcorner \phi \urcorner \,\&\, \phi(x)\big)$$

With all instances of "ϕ" replaced with ϕ. This sentence effectively says that the expression ϕ is true of its own Gödel number if ϕ contains only x free.

With all of the preliminary material out of the way, we can finally get into the proof.

Lemma 1

Let $diag$ be a function that maps $\ulcorner \phi \urcorner$ to the Gödel number of the diagonalization of ϕ. $diag$ is computable.

Proof: Let $len(n)$ be defined as the least x such that $0 < x \,\&\, n < 10^x$. The value of len is the length of the numeric representation of a number n. For instance, $len(45) = 2$ and $len(65535) = 5$. len is computable, since "less than" is computable, "the least such that" is computable, and exponentiation is computable. Then, let $m * n = m \cdot 10^{len(n)} + n$. $*$ is computable, since addition, multiplication, exponentiation, and len are all computable. $*$ then represents the concatenation of two numbers. For instance, $29 * 7889 = 297889$. Finally, let num be defined recursively: $num(0) = 6$, and $num(n + 1) = 68 * num(n)$. num then takes a number as an argument and has as its value the Gödel number of the expansion of its argument.

For instance, $num(2) = 68686$. num is then computable, since recursion is computable and $*$ is computable.

We can then define $diag(n)$ to be $145217885 * \Big(num(n) * \big(3 * (n * 2)\big)\Big)$. $diag$ is computable, since $*$ is computable and num is computable. Lemma 1 has been proven.

The diagonal lemma is what we are about to prove. It is essential to demonstrate the final result. The diagonal lemma gives the existence of certain self-referential sentences of arithmetic—sentences that are true if and only if some condition is true of that sentence. In fact, we're going to prove that these sentences exist for *all* such "conditions." Without further ado, the diagonal lemma:

Lemma 2 (The Diagonal Lemma)

In any theory in which $diag$ is representable, for all formulas ϕ with one free variable there exists a sentence ψ such that:

$$\psi \leftrightarrow \phi(\ulcorner \psi \urcorner)$$

Proof: The first thing we're going to do is create a predicate that represents $diag$ in this theory. Representability is a property of an n-ary function that entails its ability to be defined as a formula with $n + 1$ free variables. The first n free variables of this formula are just the arguments of the given function, and the $n + 1^{\text{th}}$ free variable of this formula is the result of applying the function on the first n arguments. This formula cannot merely be a predicate defined by interpretation—it has to be "embedded" into the language, hence the emphasis on computability in Lemma 1. For if a function is computable, it is representable in arithmetic. Representing $diag$ in arithmetic yields a predicate that we will call D^2. Note that D^2 is really a formula of the language of arithmetic with two free variables, not a predicate defined in an interpretation, as aforementioned. Define D^2 as such: if $diag(n) = k$ then $D^2\mathbf{nk}$. In other words, if the Gödel

number of the diagonalization of n is k, then $D^2\mathbf{nk}$. Remember than $\mathbf{n}$'s relation to n is such: if the Gödel number of an expression ϕ is n, then $\ulcorner \phi \urcorner = \mathbf{n}$.

Then we define a sentence $F(x) = (\exists y)\big(D^2xy \,\&\, \phi(y)\big)$. Let n be the Gödel number of F. Then, let $\psi = (\exists x)\Big(x = \mathbf{n} \,\&\, (\exists y)\big(D^2xy \,\&\, \phi(y)\big)\Big)$. Recall that the diagonalization of a formula is $(\exists x)\big(x = \ulcorner \phi \urcorner \,\&\, \phi(x)\big)$. Since the Gödel number of F is n, $\ulcorner F \urcorner = \mathbf{n}$. It's easy then to see that ψ is the diagonalization of F. ψ is logically equivalent to $(\exists y)\big(D^2\mathbf{n}y \,\&\, \phi(y)\big)$. So, we have:

$$\psi \leftrightarrow (\exists y)\big(D^2\mathbf{n}y \,\&\, \phi(y)\big)$$

Let k be the Gödel number of ψ. So, $\ulcorner \psi \urcorner = \mathbf{k}$. So, we have $D^2\mathbf{nk}$, from which:

$$(\forall y)(D^2\mathbf{n}y \leftrightarrow y = \mathbf{k})$$

Follows. Since $D^2\mathbf{n}y$ is logically equivalent to $y = \mathbf{k}$, $(\exists y)\big(D^2\mathbf{n}y \,\&\, \phi(y)\big)$ is logically equivalent to $(\exists y)\big(y = \mathbf{k} \,\&\, \phi(y)\big)$. And since ψ is logically equivalent to $(\exists y)\big(D^2\mathbf{n}y \,\&\, \phi(y)\big)$, we have:

$$\psi \leftrightarrow (\exists y)\big(D^2\mathbf{n}y \,\&\, \phi(y)\big)$$

Which is logically equivalent to:

$$\psi \leftrightarrow \phi(\mathbf{k})$$

And since the Gödel number of ψ is k, we have (finally):

$$\psi \leftrightarrow \phi(\ulcorner \psi \urcorner)$$

Thus, proving the diagonal lemma.

Now that we have worked through Gödel's first incompleteness theorem up through the diagonal lemma, we are ready to present the main argument. This hinges on the diagonal lemma, so the following demonstration only works for systems that are strong enough to represent $diag$, i.e. systems of arithmetic at least as strong as Robinson arithmetic. Anyway, here's the main result:

Theorem 1

In any formal system in which $diag$ is representable that contains a predicate $Know$ such that $Know(\ulcorner P \urcorner) \rightarrow P$, there exists some sentence ϕ such that ϕ & $\neg Know(\ulcorner \phi \urcorner)$.

Proof: By the diagonal lemma, we are able to demonstrate in any theory in which $diag$ is representable that for any formula ϕ, there exists sentence ψ such that:

$$\psi \leftrightarrow \phi(\ulcorner \psi \urcorner)$$

Consider the predicate $Know$, such that $Know(\ulcorner p \urcorner) \rightarrow p$. In essence, $Know$ encodes some subset of the theorems of arithmetic. There may be some way of obtaining knowledge (and there should be), so we may have some $\omega(\ulcorner p \urcorner) \rightarrow Know(\ulcorner p \urcorner)$ where ω is some sufficient condition for knowledge. The specifics of ω are left unspecified until the section on Tarski's indefinability theorem, because it generally does not have an effect on the theorem.

By the diagonal lemma, there exists some sentence ψ such that the following:

$$\vdash \psi \leftrightarrow \neg Know(\ulcorner \psi \urcorner)$$

ψ would then "read" as "I am not known." We can write this result as a first-order schema:

$$(\exists \phi)\big(\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)\big)$$

Now consider the following deduction:

{}	(1)	$(\exists \phi)\big(\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)\big)$	Lemma 1
{2}	(2)	$\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)$	Premise
{3}	(3)	$\neg \phi$	Premise
{2, 3}	(4)	$Know(\ulcorner \phi \urcorner)$	2, 3, Tautological inference
{2, 3}	(5)	ϕ	4, Definition of $Know(\ulcorner \phi \urcorner)$
{2}	(6)	$\neg \phi \rightarrow \phi$	3, 5, Conditionalization
{2}	(7)	ϕ	6, Tautological inference
{2}	(8)	$\neg Know(\ulcorner \phi \urcorner)$	2, 7, Tautological inference
{2}	(9)	ϕ & $\neg Know(\ulcorner \phi \urcorner)$	7, 8, Conjunctive addition
{2}	(10)	$(\exists \phi)\big(\phi$ & $\neg Know(\ulcorner \phi \urcorner)\big)$	9, Existential generalization
{}	(11)	$(\exists \phi)\big(\phi$ & $\neg Know(\ulcorner \phi \urcorner)\big)$	10, ES procedure

This final sentence reads that there exists a sentence ϕ that is both true and not known. Q. E. D. It's worth noting that I didn't add a dependency to $Know(\ulcorner\ \phi\ \urcorner) \rightarrow \phi$. I did this because I'm assuming that $Know(\ulcorner\ \phi\ \urcorner) \rightarrow \phi$ is some axiom of the system, and hence, needs no dependency line. If one desires to add the dependency line and conditionalize at the end, then you would simply end up with a similar theorem on your last line:

$$(Know(\ulcorner\ \phi\ \urcorner) \rightarrow \phi) \rightarrow (\exists\phi)\big(\phi\ \&\ \neg Know(\ulcorner\ \phi\ \urcorner)\big)$$

Which reads "If knowledge is factive, then there exists some sentence that is true and not known."

So, there's the proof explained. The key takeaway here is that $Know$ is some predicate represented in a theory of arithmetic at least as strong as Robinson arithmetic, $Know(\ulcorner\ P\ \urcorner) \rightarrow P$, and that there is some sentence ϕ that is true but $Know(\ulcorner\ \phi\ \urcorner)$ doesn't hold, with that sentence "saying" "I am not known."

This is not to say that there is any particular blindspot that is necessarily unknowable. Consider knowledge at two distinct points in time: $Know_A$ represents the contents of one's knowledge at time A, and $Know_B$ represents the contents of one's knowledge at time B. It's a given that the contents of knowledge change over time as one learns and adds more to their knowledge. So, we may have a proposition ϕ that is unknown due to the diagonal lemma at time A:

$$\phi\ \&\ \neg Know_A(\ulcorner\ \phi\ \urcorner)$$

Now consider the case that through some manner (the manner in which this occurs is not significant to me) ϕ is learned at some point between time A and time B. Then we will indeed have

$$Know_B(\ulcorner\ \phi\ \urcorner)$$

But the diagonal lemma necessitates that there *must be some* proposition that is unknown. So, at time B, the set of Gödel blindspots changes to include ψ:

$$\psi \,\&\, \neg Know_B(\ulcorner\, \psi \,\urcorner)$$

So Gödel blindspots always exist, but they can change depending on the specific knowledge predicate in question. There is not one specific object of knowledge that is necessarily unknown.

Tarski's Indefinability Theorem

One might have the worry that the prior result conflicts with Tarski's indefinability theorem. This section will be dedicated to resolving any concerns that the reader may have about this. This section will be rather informal (a step back from the proof in the prior section), but should still be enough to convince the reader that there is no conflict between theorem 1 and Tarski's indefinability theorem. At the end of the section, I will present an alternate proof of theorem 1 that actually shows it to be a *corollary* of Tarski's indefinability theorem.

An Informal Demonstration of Consistency

Tarski's indefinability theorem states that "the set of Gödel numbers of sentences true in N is not definable in arithmetic." This can be rewritten as the second-order principle:

$$\neg(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$$

Theorem 1 is derivable from the following sentence:

$$(\exists \phi)\big(\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)\big)$$

In effect, this states that if knowledge is factive, then there exists a sentence that is true if and only if it's not knowable. First, it's worth noting that theorem 1 is also derivable from Tarski's theorem:

{ }	(1)	$\neg(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	Tarski's theorem
{ }	(2)	$(\forall F)\neg(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	1, Quantifier exchange
{ }	(3)	$\neg(\forall \phi)\big(\phi \leftrightarrow Know(\ulcorner \phi \urcorner)\big)$	2, Universal specification
{ }	(4)	$(\exists \phi)\neg\big(\phi \leftrightarrow Know(\ulcorner \phi \urcorner)\big)$	3, Quantifier exchange
{ }	(5)	$(\exists \phi)\big(\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)\big)$	4, Replacement

Now, obviously this doesn't consider the factivity of knowledge. The very informal way I will attempt to demonstrate consistency between theorem 1 and Tarski's theorem is by first considering knowledge that is both factive and logically omniscient (all true things are known),

showing that that is *inconsistent*, then showing that the proof fails if knowledge is only factive (with some other logically non-omniscient method for knowledge acquisition).

Consider the set of sentences containing Tarski's theorem and a concept of knowledge that holds knowledge as factive and that all truths are known. Now consider the following trivial deduction:

{}	(1)	$\neg(\exists F)(\forall \phi)\big(p \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	Tarski's theorem
{}	(2)	$(\forall \phi)(Know(\ulcorner \phi \urcorner) \leftrightarrow \phi)$	Definition of knowledge
{}	(3)	$Know(\ulcorner \psi \urcorner) \leftrightarrow \psi$	2, Universal specification
{}	(4)	$\psi \leftrightarrow Know(\ulcorner \psi \urcorner)$	3, Tautological inference
{}	(5)	$(\forall \phi)\big(\phi \leftrightarrow Know(\ulcorner \phi \urcorner)\big)$	4, Universal generalization
{}	(6)	$(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	5, Existential generalization
{}	(7)	$\bot$	1, 6, $\bot$ introduction

Now consider the sister deduction, without the assumption that all truths are known:

{}	(1)	$\neg(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	Tarski's theorem
{}	(2)	$(\forall \phi)(Know(\ulcorner \phi \urcorner) \rightarrow \phi)$	Definition of knowledge
{}	(3)	$Know(\ulcorner \psi \urcorner) \rightarrow \psi$	2, Universal specification
{}	(4)	$\psi \leftarrow Know(\ulcorner \psi \urcorner)$	3, Tautological inference
{}	(5)	$(\forall \phi)\big(\phi \leftarrow Know(\ulcorner \phi \urcorner)\big)$	4, Universal generalization
{}	(6)	$(\exists F)(\forall \phi)\big(\phi \leftarrow F(\ulcorner \phi \urcorner)\big)$	5, Existential generalization

Without the assumption that all truths are known, it lacks the strength to be in contradiction to Tarski's indefinability theorem. Intuitively, we can think of it like this: Tarski's indefinability theorem states that there is no Gödel representation of the theorems of arithmetic in arithmetic. The reason this doesn't cause issues for theorem 1 is because, without the assumption that all truths are known, only a single direction of the mutual implication is met. Knowledge in this case embeds a *subset* of the theorems of arithmetic. In fact, Tarski's theorem can be thought of as another way to demonstrate theorem 1, because knowledge, if represented in arithmetic, *cannot* be

comprehensive of all truths, or else it will be in contradiction with Tarski's theorem, since $Know$ would then just be an alias for the truth predicate.

This alternate way to prove theorem 1 is as follows:

Theorem 1 (Proven Another Way)

In any formal system that represents $diag$ and a predicate $Know$ such that $Know(\ulcorner \phi \urcorner) \rightarrow \phi$, there exists some sentence ϕ such that ϕ & $\neg Know(\ulcorner \phi \urcorner)$.

Proof: Any formal system that represents $diag$ is strong enough to satisfy the conditions for Tarski's indefinability theorem. Tarski's theorem states the following:

$$\neg(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$$

So, consider the following deduction:

{}	(1)	$\neg(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	Tarski's theorem
{2}	(2)	$(\forall \phi)(Know(\ulcorner \phi \urcorner) \rightarrow \phi)$	Knowledge factivity
{3}	(3)	$(\forall \phi)\big(\phi \rightarrow Know(\ulcorner \phi \urcorner)\big)$	Logical omniscience
{2}	(4)	$Know(\ulcorner \phi \urcorner) \rightarrow \phi$	2, Universal specification
{3}	(5)	$\phi \rightarrow Know(\ulcorner \phi \urcorner)$	3, Universal specification
{2, 3}	(6)	$\phi \leftrightarrow Know(\ulcorner \phi \urcorner)$	4, 5, ↔ addition
{2, 3}	(7)	$(\forall \phi)\big(\phi \leftrightarrow Know(\ulcorner \phi \urcorner)\big)$	6, Universal generalization
{2, 3}	(8)	$(\exists F)(\forall \phi)\big(\phi \leftrightarrow F(\ulcorner \phi \urcorner)\big)$	7, Existential generalization
{2, 3}	(9)	$\bot$	1, 8, ⊥ introduction
{2}	(10)	$\neg(\forall \phi)\big(\phi \rightarrow Know(\ulcorner \phi \urcorner)\big)$	3, 9, Reductio ad absurdum
{2}	(11)	$(\exists \phi)\neg\big(\phi \rightarrow Know(\ulcorner \phi \urcorner)\big)$	10, Quantifier equivalence
{2}	(12)	$(\exists \phi)\big(\phi \,\&\, \neg Know(\ulcorner \phi \urcorner)\big)$	11, Replacement

The result on line 12 is just theorem 1, this time with a more explicit reference to the premise that knowledge is factive. So, theorem 1 is proven again, this time as a corollary to Tarski's indefinability theorem. It's worth noting that in the RAA step on line 10, one could've instead performed RAA on premise 2, rejecting factivity of knowledge. This is another result, and instead says that, if all true things are known, then one also knows some falsities. This is another interesting

result that may be used to argue against factivity, but in general, lots of contemporary epistemologists seem to accept factivity, so in my opinion, it makes more sense to assert that there always exist unknown sentences (Mitova 2018).

Defining Gödel Blindspots

Now I would like to give an account of Gödel blindspots, the kind of blindspots demonstrated to exist by theorem 1. Before we establish what Gödel blindspots are, we need a generalized definition of epistemic blindspots, so that we can determine what the difference is.

Sorensen defines epistemic blindspots as "propositions that are inaccessible through weak constraints," where he defines the weakest "constraint" as being logic. I prefer a more descriptive definition: true but unknowable propositions. That is, a p such that $p \,\&\, \neg\Diamond Kp$. One set of such propositions consists of propositions with form $p \,\&\, \neg Kp$ (under the assumption that knowledge distributes over conjunction). The proof is as follows:

{1}	(1)	$p \,\&\, \neg Kp$	Premise
{2}	(2)	$K(p \,\&\, \neg Kp)$	Premise
{2}	(3)	$Kp \,\&\, K\neg Kp$	2, Distribution
{2}	(4)	$K\neg Kp$	3, Conjunctive simplification
{2}	(5)	$\neg Kp$	4, Factivity of knowledge
{2}	(6)	Kp	3, Conjunctive simplification
{2}	(7)	$\bot$	5, 6, $\bot$ introduction
{ }	(8)	$\neg K(p \,\&\, \neg Kp)$	2, 7, Reductio ad absurdum
{ }	(9)	$\Box\neg K(p \,\&\, \neg Kp)$	8, Necessitation
{ }	(10)	$\neg\Diamond K(p \,\&\, \neg Kp)$	9, Modal equivalence
{1}	(11)	$(p \,\&\, \neg Kp) \,\&\, \neg\Diamond K(p \,\&\, \neg Kp)$	1, 10, Conjunctive addition

As an aside, the proof bears much resemblance to the deduction in Fitch's paradox. As is seen, the final proposition has the form $p \,\&\, \neg\Diamond Kp$, meaning $p \,\&\, \neg Kp$ is an epistemic blindspot.

How does this tie into theorem 1? As we demonstrated earlier, we proved something to the effect of $(\exists\phi)\big(\phi \,\&\, \neg Know(\ulcorner\phi\urcorner)\big)$, which has a form similar to the modal sentence $(\exists p)(p \,\&\, \neg Kp)$. Here we're going to depart from our general definition—as you can see, an instantiation of the existential quantifier does *not* yield a sentence of the form $p \,\&\, \neg\Diamond Kp$. In fact, that result is impossible. The reason is that necessitating our theorem yields $\Box(\exists p)(p \,\&\, \neg Kp)$. As

Plantinga notes, you cannot get $(\exists p)\Box(p \,\&\, \neg Kp)$ from this sentence, which is necessary to get $(\exists p)(p \,\&\, \neg\Diamond Kp)$, a sentence that expresses the existence of general blindspots. (Plantinga 1974).

So, Gödel blindspots are not a subset of epistemic blindspots. What are they, then? This is the account I will give: Gödel blindspots are unknown (note: not *unknowable*) propositions that are unknown due to the diagonal lemma. A particular Gödel blindspot can come to be known by augmenting the knowledge representation to include that proposition (such as knowledge acquisition over time), but the diagonal lemma says that there will always be some propositions that are not known, no matter how much the knowledge representation is augmented to include arithmetic truths (for, if there was such a representation of knowledge, it would come in conflict with Tarski's indefinability theorem). Those propositions that are not known due to the diagonal lemma but are true are Gödel blindspots.

Philosophical Reflections

In this section of the paper, I will start with a discussion of Fitch's paradox of knowability, which is a seemingly absurd deduction that stems from a claim of anti-realism to derive epistemic trivialism (that all truths are known). I will give a defense of rejecting the premise of Fitch's paradox that all truths are knowable under the conditions that "knows that" is a predicate represented in arithmetic, and that the term "truths" is restricted to "truths in arithmetic." Then, we will take a break from Fitch's paradox to see how we can strengthen theorem 1 using some natural language arguments, attempting to break away from these arithmetic restrictions and formalize general blindspots (propositions that are true and *unknowable*, not merely unknown). Finally, we will revisit Fitch's paradox with our strengthened results, and reject the main premise of Fitch's paradox with an even stronger argument.

Gödel Blindspots and the K Principle

Fitch's paradox of knowability is a deduction in alethic-epistemic logic that takes the following premise known as the K Principle:

$$(\forall p)(p \rightarrow \Diamond Kp)$$

Stating that all truths are knowable, and derives the following conclusion:

$$(\forall p)(p \rightarrow Kp)$$

Stating that all truths are known. The only assumptions made about knowledge in the derivation are that knowledge is factive ($Kp \rightarrow p$) and that knowledge distributes over conjunction ($K(p \,\&\, q) \rightarrow Kp \,\&\, Kq$). As one of my professors says, "one man's modus ponens is another man's modus tollens." There are three general routes for responding to this result:

1. Reject the K Principle and embrace the existence of blindspots
2. Accept the conclusion, and embrace the notion that all truths are known

3. Reject the deduction as a whole

At this point, I would like to give a defense of choosing 1 given certain assumptions. As we've demonstrated through theorem 1, there exist propositions that are both true and not known. If expressed modally, we get:

$$(\exists p)(p \,\&\, \neg Kp)$$

Which is in direct contradiction with the conclusion from Fitch's paradox. However, we must not forget the premises that made to get to this conclusion. Theorem 1 is a theorem about knowability when represented in arithmetic—not knowability *in univsersum*. This leads us to the more general question—is there a good reason for wanting to represent knowability in arithmetic? Since this question is deserving of its own paper, I'll leave my response brief: there are reasons for wanting to represent knowledge as an arithmetic predicate, and I omit those reasons for the sake of brevity. However, under the circumstance where we *do* want to represent knowledge in arithmetic, it's clear that we must accept theorem 1, which is in direct contradiction to the conclusion of Fitch's paradox. This means that we certainly should not accept 2 as our solution of Fitch's paradox. Therefore, if we accept the deduction, then obviously that omits 3 as an option, and leaves us with option 1.

There are reasons for preferring 3. If we continue to stick with the theme of "representability in arithmetic," we must reject alethic modal operators, since necessity cannot be represented in arithmetic, and possibility is defined in terms of necessity. The reason necessity cannot be represented in arithmetic is simple: $(\forall p)(p \leftrightarrow \Box p)$ is consistent in modal logic (it restricts the frame to a single reflexive world), but $(\forall \phi)\big(\phi \leftrightarrow L(\ulcorner \phi \urcorner)\big)$ (where L is the representation of $\Box$) is inconsistent due to Tarski's indefinability theorem, since L would then just

contain the theorems of arithmetic. So, there is no general representation of □ as a predicate in arithmetic. Therefore, since the K Principle can't even be represented, the deduction is bogus.

A solution is to simply use quantified modal logic as the basis for arithmetic and use raw modal operators instead of representing them in arithmetic. However, Quine has famously argued against QML due to certain ontological obligations it brings with it (Quine 1947). While these arguments aren't as problematic as they once were (Marcus 1995), they are still worth considering. If, despite this, one uses quantified modal logic as the basis for arithmetic and decides to represent knowledge in arithmetic, then I see no reason why they would accept any other conclusion other than the falsity of the K Principle *in this application*.

Strengthening Theorem 1 with Natural Language

At this point, I would like to expand into more general forms of epistemology—using Gödel blindspots to reject the K Principle in applications other than representability in mathematics. First, it's worth mentioning structuralism, the notion that the universe is a structure in mathematics. Under some interpretations of structuralism, knowledge would then be a structure in mathematics, which could then fall victim to theorem 1. However, structuralism is very broad and ununified (Schmidt and Heinz-Juergen 2019), in addition to the fact that I'm not well read enough on structuralism to make definitive judgements about the position, so it's not possible to make a generalizing statement such as "A structuralist should believe the result from theorem 1 and apply it to general epistemology." What we can, do, however, is identify the natural language equivalent to the Gödel-like sentence, the ϕ such that $\phi \leftrightarrow \neg Know(\ulcorner \phi \urcorner)$.

This natural language equivalent is simply how this Gödel-like sentence is read: "I am not known." The difficulty in establishing theorem 1 was a proof that such a sentence existed in arithmetic, but we know such a sentence exists in natural language because we can state it in its

entirety—it exists on the paper. We can formalize the sentence in modal form in the same manner that we've been doing throughout the paper:

$$p \leftrightarrow \neg Kp$$

Through the same basic proof give to establish theorem 1, we can establish:

$$p \,\&\, \neg Kp$$

Which, then, existentially generalizing to avoid assigning p:

$$(\exists p)(p \,\&\, \neg Kp)$$

Conjecture Ka

This demonstrates the existence of epistemic blindspots, but this time without the restriction that discussion be limited to representable predicates in arithmetic, and that p be an arithmetic sentence. We simply used theorem 1 as an inspiration to find such a sentence in natural language.

Strengthening Further

Is it possible to obtain a general blindspot using this procedure? To come up with a self-referential sentence that is *necessarily* not known (or equivalently, not possible to know)? Such a sentence should look along the lines of the p such that:

$$\Box(p \leftrightarrow \neg Kp)$$

From here, we can do a line-by-line deduction to demonstrate *conjecture Kb*:

{1}	(1)	$(\exists p)\Box(p \leftrightarrow \neg Kp)$	Premise
{2}	(2)	$\Box(q \leftrightarrow \neg Kq)$	Premise for ES
{3}	(3)	$q \leftrightarrow \neg Kq$	Premise
{3}	(4)	$q \,\&\, \neg Kq$	Proof technique used in theorem 1
{}	(5)	$(q \leftrightarrow \neg Kq) \rightarrow (q \,\&\, \neg Kq)$	2, 3, Conditionalization
{}	(6)	$\Box\big((q \leftrightarrow \neg Kq) \rightarrow (q \,\&\, \neg Kq)\big)$	4, Necessitation
{}	(7)	$\Box(q \leftrightarrow \neg Kq) \rightarrow \Box(q \,\&\, \neg Kq)$	5, Distribution
{2}	(8)	$\Box(q \,\&\, \neg Kq)$	2, 6, Modus ponens
{2}	(9)	$\Box q \,\&\, \Box\neg Kq$	7, Distribution over conjunction

{2}	(10)	$\Box q$	9, Conjunctive simplification
{2}	(11)	q	10, Axiom M
{2}	(12)	$\Box \neg Kq$	9, Conjunctive simplification
{2}	(13)	$\neg \Diamond Kq$	12, Modal equivalence
{2}	(14)	$q \ \& \neg \Diamond Kq$	11, 13, Conjunctive addition
{2}	(15)	$(\exists p)(p \ \& \neg \Diamond Kp)$	14, Existential generalization
{1}	(16)	$(\exists p)(p \ \& \neg \Diamond Kp)$	1, 2, 15, ES procedure

So there exists a p that is a traditional blindspot, true and not possible to know. The reason Gödel blindspots are merely not known instead of not possible to know is because there is no way to guarantee that p has the same truth value in each possible world. That would require necessitating inside the existential quantifier in theorem 1, which is not possible, as Plantinga states (Plantinga 1974). For theorem 1, we prove that some blindspot necessarily exists in every possible world, but not that a sentence exists such that it's not known in any possible world. However, we can do such an action here, because instead we know that there exists a sentence that says "Necessarily, I am not known."

It's worth noting that we make use of axiom M in the deduction, that $\Box p \rightarrow p$. We haven't made use of this axiom in the past (we've just been making use of the modal system K), but most philosophers think that M is true, since what is necessary should also be what's true (Garson 2013).

Factive Predicates, Fitch's Paradox Revisited, and Anti-Realism

"Knowledge" is a very strong word. However, in all of our demonstrations, the only feature of knowledge we ever actually *used* was factivity. In this case, what separates "knowledge" from "true belief?" In fact, all of our demonstrations apply just as well for true belief. In this case, we not only have epistemic blindspots that are true but not possible to know, but we, in fact, have epistemic blindspots that are true and are not possible to have true belief in! In fact, the same argument applies for any other factive predicate, such as "demonstrability," assuming you can

demonstrate something only if it's true. Then, we have sentences like "Necessarily, I am not demonstrated," which are true but not possible to demonstrate.

At this point, we have a good place to re-enter our discussion about Fitch's paradox. Previously, we could only object to Fitch's paradox on the basis of rejecting the K Principle by virtue of a modus tollens argument, and only when we consider knowledge as a predicate represented in arithmetic. Now, we have the tools to reject the K Principle directly with conjecture Kb, and without the representability restriction on the knowledge operator. Recall that the K principle states that:

$$(\forall p)(p \rightarrow \Diamond Kp)$$

And that conjecture Kb states that:

$$(\exists p)(p \,\&\, \neg\Diamond Kp)$$

Which are contradictory. Since conjecture Kb was proven using an instantiation of a sentence about natural language (the sentence "Necessarily, I am not known"), we have a good basis to reject the K principle on natural language arguments alone.

Furthermore, we can reject other assertions with the form $(\forall p)(p \rightarrow \Diamond Fp)$ where F is any factive modal operator. We can reject the assertion that "all truths are demonstrable," that "all truths are able to be the contents of true belief," and similar assertions involving all other factive predicates, since we have given good evidence to support the more general statement (where $(\forall F)$ quantifies over the strong modal operator associated with all accessibility relations):

$$(\forall F)(\forall p)\big((Fp \rightarrow p) \rightarrow (\exists p)(p \,\&\, \neg\Diamond Fp)\big)$$

By creating a sentence p such that, for some F:

$$\Box(p \leftrightarrow \neg Fp)$$

And existentially quantifying into the box, to yield:

$$(\exists p)\Box(p \leftrightarrow \neg Fp)$$

Which then gives support for conjecture Kb, with the sentence p being logically equivalent to the sentence "necessarily, I am not F" for any factive predicate F.

Even though we reject the K Principle, this is not a strong enough argument to reject anti-realism (the position that truth is epistemic) entirely. As Hand suggests, anti-realism doesn't necessarily imply the K principle (Hand 2003). However, even Hand admits:

> "In extraordinary cases, where interference is produced by attempts to perform verification procedures for propositions about the performance of verification procedures, *truth can hide*. [emphasis added] [… T]he antirealist has no difficulty in attributing an understanding of p & $\neg Kp$, a grasp of what the *truth* of p & $\neg Kp$ consists in. It is precisely in virtue of this understanding that a competent user can see immediately that if p & $\neg Kp$ is true, there is no way to verify it. Now, everything just said is available to the antirealist, so the antirealist can give an account of the knowability paradox on which the truth of p & $\neg Kp$ cannot be discovered.

This seems to suggest the existence of epistemic blindspots (though he does not use this term), but that these blindspots must be explicitly crafted and are not "naturally occurring," so to speak. One instance is p & $\neg Kp$, but another could possibly be the p such that $\Box(p \leftrightarrow \neg Kp)$. Because the antirealist understands what it means for that sentence to be true, the antirealist can give an account of these blindspots. However, their truth cannot be discovered, because such truths can hide due to "interference."

So, while I don't make an argument against anti-realism as a whole, I do make an argument against the K principle. The K Principle directly contradicts the existence of blindspots like

"Necessarily, I am not known," a blindspot inspired by the sentence proven by the diagonal lemma: $(\exists p)\big(p \leftrightarrow \neg Know(\ulcorner\, p\, \urcorner)\big)$, of which an instantiation of p would be a blindspot.

Some Notes on Symbolizations of Self-Referential Propositions

This entire chapter, I have been under the assumption that "Necessarily, I am P" for some modality P can be symbolized as: the p such that $\Box(p \leftrightarrow Pp)$. However, this could also be symbolized in other, more bizarre ways, depending on whether or not the word "necessarily" is taken to belong to the part of the sentence contained in the self-reference. This is made clearer by asking the question: when we say "I" in a sentence, what do we refer to? In this case, is the word "necessarily" included in the scope of the word "I?" If so, then we would symbolize it differently than I have been in this essay. We can force a symbolization of $\Box(p \leftrightarrow Pp)$ by saying: the sentence $p =$ "I am P" holds in every possible world. Working the necessary modality into the sentence itself is tricky, because it might be encapsulated by the self-reference.

I stand by the interpretation that the word "necessarily" is *not* contained by the word "I," but there is a strong case to be made against this. I will not be making an argument for my position in this paper, though, because that is getting out of the scope of what this essay is. I believe it is worth mentioning, however.

Bibliography

Sorensen, Roy. 1988. *Blindspots*. Oxford: Oxford University Press. Print.

G. Boolos, R. Jeffrey. 1989. *Computability and Logic* (3 ed). Cambridge: Cambridge University Press. Print.

Plantinga, Alvin. 1974. *The Nature of Necessity*. Oxford: Oxford University Press. Print.

Mates, Benson. 1972. *Elementary Logic* (2 ed). Oxford: Oxford University Press. Print.

Brogaard, Berid, and Salerno. Fitch's Paradox of Knowability. *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). ed. Edward Zalta. https://plato.stanford.edu/entries/fitch-paradox/.

Schmidt, Heinz-Juergen. Structuralism in Physics. *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). ed. Edward Zalta. https://plato.stanford.edu/archives/win2019/entries/physics-structuralism.

Quine, W. V. The Problem of Interpreting Modal Logic. *The Journal of Symbolic Logic*: 12 (1947): 43-48. https://doi.org/10.2307/2267247.

Marcus, Ruth. 1995. A Backwards Look at Quine's Animadversions on Modalities. In *Perspectives on Quine*: 230-243. ed. R. Barrett, R. Gibson.

Hand, Michael. Knowability and Epistemic Truth. *Australasian Journal of Philosophy*: 81 (2003): 216-228.

Humberstone, Lloyd. Supervenience, Dependence, Disjunction. *Logic and Logical Philosophy*: 28 (2019). http://dx.doi.org/10.12775/LLP.2018.007.

Garson, James. 2013. *Modal Logic for Philosophers* (2 ed). Cambridge: Cambridge University Press. Print.

Mitova, Veli. 2018. *The Factive Turn in Epistemology.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316818992.

Conclusion

I have proven that, when knowledge is represented in arithmetic, there exist self-referential blindspots. These blindspots are called Gödel blindspots, and are not quite a subset of traditional blindspots, but I call them blindspots due to the fact that they must exist. Gödel blindspots become Gödel blindspots by virtue of being unknowable due to the diagonal lemma. Such self-referential blindspots may be read in natural language as "I am not known." By virtue of us being able to say that sentence in natural language, I conjecture that self-referential blindspots exist through natural language, if one strengthens the sentence to say, "Necessarily, I am not known." I conjecture this because I have not given a strong defense of this position aside from "we can say such a sentence." Obviously, an argument can be made against the sentence from a position of absurdity or nonsensicalness.

One can derive from the sentence "Necessarily, I am not known," the following: $(\exists p)(p \,\&\, \neg\Diamond Kp)$, which is in direct contradiction to the K Principle, $(\forall p)(p \rightarrow \Diamond Kp)$, used to derive epistemic trivialism in Fitch's paradox of knowability. Such blindspots also exist for any other factive predicate, such as true belief ("Necessarily, I am not the contents of true belief" is impossible to be the contents of true belief) and demonstrability ("Necessarily, I am not demonstrated" is impossible to be demonstrated). This goes to reject any other K Principle analogue with the form $(\forall p)(Fp \rightarrow p)$ & $(\forall p)(p \rightarrow \Diamond Fp)$ where F is some modal operator.

This paper did not inspect constructivist responses to self-referential blindspots, not did it analyze anti-realism specifically instead of merely investigating the K Principle. These endeavors are left to a future work. However, I feel strongly about rejecting the K principle in classical logic, since I have provided yet another example of an epistemic blindspot in addition to the well-known $p \,\&\, \neg Kp$, that being the p such that $\Box(p \leftrightarrow \neg Kp)$. There are many other types of examples, but I

believe self-referential blindspots are especially interesting, since they deny their own knowability. Self-referential blindspots are unlikely to greatly impact day-to-day life, unlike perhaps some of the blindspots mentioned in Sorensen's book. However, I believe this has been sufficiently interesting to warrant a study.