KK, Knowledge, Knowability

Weng Kin San

University of Southern California, USA
wsan@usc.edu

Abstract

KK states that knowing entails knowing that one knows and K¬K states that not knowing entails knowing that one does not know. In light of the arguments against KK and K¬K, one might consider modally qualified variants of those principles. According to WEAK KK, knowing entails the *possibility* of knowing that one knows. And according to WEAK K¬K, not knowing entails the *possibility* of knowing that one does not know. This paper shows that WEAK KK and WEAK K¬K are much stronger than they initially appear. Jointly, they entail KK and K¬K. And they are susceptible to variants of the standard arguments against KK and K¬K. This has interesting implications for the debate on positive introspection and for deeper issues concerning the structure and limits of knowability.

1 Introduction

According to the KK principle, anything known is known to be known:

(KK) If one knows that φ, then one knows that one knows that φ ($K\varphi \rightarrow KK\varphi$).[1]

Defenders of KK include Plato, Aristotle, St. Augustine, Aquinas, Locke, and in more recent history, Hintikka (1962) and Stalnaker (2006).[2] Despite its impressive pedigree, KK has since fallen into disrepute, largely due to Williamson's (2000) influential margin-of-error argument. Nevertheless, recent years have seen attempts at rehabilitating KK or variants of it.[3] While KK remains contested, there is near unanimous consensus against its negative counterpart, according to which anything not known is known not to be known:[4]

[1] This and later principles are really quantified ones ranging over all agents, propositions, and so on. Such quantification is henceforth always left implicit.

[2] See Hintikka (1962, ch. 5.3) for a more comprehensive overview of the history of KK.

[3] See Greco (2014a,b, 2016), Goodman & Salow (2018), Das & Salow (2018), and Dorst (2019).

[4] As Holliday puts it, 'the rejection of…5 [K¬K] is universal among epistemologists' (Holliday, 2018, 362). Indeed, even most defenders of KK reject K¬K (e.g. Hintikka (1962) and Stalnaker (2006)).

(K¬K) If one does not know that φ, then one knows that one does not know that φ ($\neg K\varphi \rightarrow K\neg K\varphi$).

Although KK is disputed and K¬K universally rejected, standard arguments against them might appear to leave open the tenability of modally qualified variants of those principles:

(WEAK KK) If one knows that φ, then *possibly* one knows that one knows that φ ($K\varphi \rightarrow \Diamond KK\varphi$);

(WEAK K¬K) If one does not know that φ, then *possibly* one knows that one does not know that φ ($\neg K\varphi \rightarrow \Diamond K\neg K\varphi$),

where the possibility operator $\Diamond$ admits of various interpretations (metaphysical, epistemic, doxastic, evidential, deontic, and so on).[5] Much of what follows will not hang on a particular interpretation, in which case I remain neutral and simply use 'possible' and 'knowable' as catch-all terms for $\Diamond$ and $\Diamond K$, respectively.[6] However, for concreteness, I sometimes take the metaphysical reading to be canonical.

While the literature on KK, K¬K, and their relation is vast and continues to burgeon, their 'weak' counterparts have not been the subject of any similarly thorough investigation. The aim of this paper is to fill this gap and to provide a clearer view of the the epistemological landscape when it comes to principles governing knowledge and knowability. Our investigation reveals that WEAK KK and WEAK K¬K are much stronger than they initially appear and that they are susceptible to variants of standard arguments against KK and K¬K.

In particular, it turns out that WEAK KK and WEAK K¬K jointly entail KK and K¬K. Although WEAK KK and WEAK K¬K appear weak, together they have the full logical strength of their unqualified counterparts (§3). This surprising result raises questions about the individual strength of WEAK KK and WEAK K¬K. I show that although WEAK KK is genuinely logically weaker than KK, it is susceptible to a variant of the influential margin-of-error argument against KK (§4). Similarly, although WEAK K¬K is genuinely logically weaker than K¬K, it is susceptible to a variant of the influential anti-skeptical argument against K¬K (§5). Our investigation into WEAK KK and WEAK K¬K also has interesting implications for the debate on positive introspection as well as for deeper issues concerning the structure and limits of knowability (§6). But first, let us survey some initial motivations for WEAK KK and WEAK K¬K.

[5] Henceforth, quotation marks are omitted where no risk of use/mention confusion arises.

[6] I use 'it is knowable that' semi-technically as shorthand for 'it is possible that one knows that'. Alternative formalisations of an intuitive notion of knowability (many of which arise in response to Fitch's Paradox)—for instance, see Edgington (1985, 2010), Fara (2010), van Ditmarsch *et al.* (2012), and Spencer (2017)—are set aside for the purposes of this paper. Readers who deny that there is any natural reading of 'knowable' that corresponds to the possibility of knowledge are invited to henceforth substitute 'possibly known' and 'possibility of knowledge' for each occurrence of 'knowable' and 'knowability'.

2 Motivations

The key observation of this paper is that WEAK KK and WEAK K¬K together have unacceptably strong consequences. This would be an idle observation if those principles were implausible or of little interest to begin with. But that's not the case. For one, standard arguments against KK and K¬K trade on features like limitations in cognition or perception, lack of the concept of knowledge or higher-order beliefs, an uncooperative epistemic environment, and so on. On a sufficiently permissive notion of possibility (like metaphysical possibility), these aren't *necessary* features of the world. There are thus possible worlds in which the common pathologies that underlie failures of KK and K¬K are absent. So, some version of WEAK KK and WEAK K¬K seem like natural fallbacks to retreat to in order to evade the standard arguments against KK and K¬K.

Furthermore, on various other interpretations besides the metaphysical, WEAK KK and WEAK K¬K remain intuitively plausible even in light of the standard arguments against their unqualified counterparts. For instance, under the deontic interpretation of $\Diamond$ as 'it is permissible that', WEAK KK and WEAK K¬K amount to the not-obviously-objectionable principle that self-knowledge about one's knowledge or lack thereof is always permissible. Of course, not all versions of WEAK KK and WEAK K¬K are of interest. Under the temporal reading of $\Diamond$ as 'at a later time', WEAK KK and WEAK K¬K are immediately refuted by the fact that some truths about what's known or unknown will simply never be known. Overly restrictive interpretations of possibility won't yield plausible versions of those principles. But to motivate interest in the subsequent discussion, what's important is that there are variants of WEAK KK and WEAK K¬K that aren't non-starters.

Still, even if those principles are individually plausible, the observation that they are jointly too strong would remain of limited interest if the reasons for accepting one were immediate reasons for rejecting the other. But again, that's not obvious. In fact, some arguments that have been put forth in favour of WEAK KK seem to also provide at least some motivation for their negative counterparts.

One such argument is based on Moorean assertions. Consider the knowledge norm of assertion, according to which it is permissible to assert only what one knows. Part of its appeal is that it easily explains the infelicity of assertions like:

(1) # It is raining. But I don't know that it is.

Since one cannot know conjunctions of the form *φ and one doesn't know φ*, such conjunctions are unassertable by the lights of the knowledge norm.[7]

A selling point of the knowledge norm is the simple explanation it provides for the infelicity of assertions like (1). Given that, it would be desirable for the explanation to generalise to other Moorean assertions, such as:

[7] See Williamson (1996; 2000, ch.11).

(2) # It is raining. But I don't know whether I know that it is.

The dubiousness of (2) might be made vivid by the following exchange:

> **A**: How's the weather?
> **B**: It's raining.
> **A** (to **C**): **B** knows that it's raining.
> **B**: Whoa, I didn't say that! Actually, I don't know whether I know that it's raining.
> **A**: Wait...so is it raining or not?
> **B**: It is.

The explanation for the infelicity of (1) was that it is unknowable and thus not assertable without running afoul of the knowledge norm. The analogous explanation for the infelicity of (2) would be that propositions of the form *φ but I don't know whether I know that φ* are similarly unknowable. That would imply that anything of the form $K\varphi \wedge K\neg KK\varphi$ is a contradiction. In other words, the following principle holds universally:

(EPIST KK) If one knows that φ, then for all one knows, one knows that one knows that φ ($K\varphi \rightarrow MKK\varphi$).

This is the epistemic variant of WEAK KK where $\Diamond$ is interpreted as $M = \neg K\neg$ ('for all one knows'). So, proponents of the knowledge norm can straightforwardly extend their explanatory strategy for (1) and avail themselves to a simple explanation for the infelicity of (2)—so long as they posit EPIST KK. This provides some motivation for EPIST KK. Or, at least so goes the *Moorean argument* for EPIST KK.[8]

This argument is far from decisive. Not least because there are other ways of explaining the infelicity of (2) without appealing to EPIST KK.[9] But setting potential objections aside, the important observation for current purposes is that a parallel argument seems to motivate the negative counterpart of EPIST KK. Consider:

(3) # I won't take a stand on whether it's raining. But for all I know, I know that it is.[10]

The infelicity of (3) is made vivid by the following exchange:

[8] See Greco (2014b, 2015). While Greco takes the argument to motivate KK, it at most motivates EPIST KK, since the inconsistency of $K(\varphi \wedge \neg K\varphi)$ is equivalent to EPIST KK holding (thanks to Simon Goldstein for this observation).

[9] See Williamson (2013b).

[10] 'I won't take a stand on whether φ' is the verbal equivalent of a shrug to the question 'φ or not φ?'. It expresses a refusal to assent to either φ or $\neg\varphi$, which is not the same as asserting that one neither knows φ nor $\neg\varphi$.

A: Is it raining?
B (shrugs): I'd rather not take a stand on that.
A (to **C**): **B** doesn't know whether it's raining.
B: Whoa, I didn't say that! Actually, for all I know, I know that it's raining.
A: Wait…so is it raining or not?
B: I already said I'd rather not take a stand on that!

Ignorance about whether φ and a refusal to commit either way seem to make it thereby impermissible to also assert that for all one knows, one knows that φ. Extending the previous explanatory strategy, the reason must be that one can't know that *for all one knows, one knows that* φ whenever one is ignorant about whether φ. This amounts to the validity of the epistemic variant of WEAK K$\neg$K:

(EPIST K$\neg$K) If one does not know that φ, then for all one knows, one knows that one does not know that φ ($\neg K\varphi \rightarrow MK\neg K\varphi$).

For, a counterexample to EPIST K$\neg$K would imply that being ignorant about whether φ ($\neg K\varphi \wedge \neg K\neg\varphi$) can in fact be consistent with knowing that for all one knows, one knows that φ ($K\neg K\neg K\varphi$).[11]

In short, according to the argument from Moorean assertions, proponents of the knowledge norm have reason to accept EPIST KK. That's because doing so makes available to them a simple explanation of the problem with assertions like (2). But if so, then they also have some reason to accept EPIST K$\neg$K because it is what's required for the parallel explanation of the problem with assertions like (3).[12]

Another argument for WEAK KK, based on common knowledge, also seems to provide *prima facie* motivation for WEAK K$\neg$K. First, consider:

> To deactivate a bomb, members of a task force stationed at different locations must each enter the same deactivation code into the terminals at their respective locations. Once one of them enters a code, others failing to also enter the correct code into their terminals within a time limit will set off the bomb. Contrast:
>
> **GROUP.** Headquarters has just figured out the code. They send the code to the entire task force as a group through the open channel. Knowing that everyone is closely monitoring communications,

[11] For, suppose EPIST K$\neg$K fails so that $\neg K\varphi \wedge K\neg K\neg K\varphi$ for some φ. The second conjunct entails that $\neg K\neg K\varphi$. In which case, it must also be that $\neg K\neg\varphi$ (for if $K\neg\varphi$, then by the known factivity of knowledge, $K\neg K\varphi$). Therefore, it follows that one is ignorant about whether φ but also knows that for all one knows, one knows that φ ($\neg K\varphi \wedge \neg K\neg\varphi \wedge K\neg K\neg K\varphi$).

[12] Smithies (2012) uses the dubiousness of sentences similar to (3) to argue for the justificatory analogue of K$\neg$K, according to which if there is no justification for believing φ, then there is justification for believing that there is no justification for believing φ.

each member of the task force reasonably expects the others to have seen the message. They all enter the code, successfully deactivating the bomb.

PRIVATE. Through a lack of foresight, headquarters sends the code to every member of the task force separately as private messages instead. Since a hack or any other kind of suspicious activity is highly unlikely, each member of the task force reasonably assumes that the message was sent privately only to them by mistake. Headquarters must have intended to send the code to the entire task force as a group but mistakenly sent it only to them. So, each fairly but falsely infers that only they know the code. Each holds off on entering the code and the bomb isn't deactivated.

What explains the success in coordination in one case but not the other? The difference can't be knowledge of the code. In both cases, everyone knows the code. Instead, the explanation is that in the second case, it's not the case that everyone knows that everyone knows the code. Only in the first case is there not only mutual knowledge (everyone knows the code) but also second-order mutual knowledge (everyone knows that everyone knows the code), third-order mutual knowledge (everyone knows that everyone knows that everyone knows the code), and so on. Where there is n-th order mutual knowledge for all $n > 0$, there is *common knowledge*. Examples like this are supposed to demonstrate the possibility of common knowledge and the role they play in coordination.[13]

But what best explains how common knowledge is possible? According to the *common-knowledge argument*, it is some form of WEAK KK.[14] Consider the simple case involving a group consisting of just you and me. Suppose that our epistemic situation with respect to φ is symmetrical and knowably so. Roughly, symmetry means that whatever I know about φ you know too, and vice versa. And similarly, whatever is knowable to me about φ is knowable to you too, and vice versa. Then, assuming that I know that φ:

(1)	I know that φ.	assumption
(2)	You know that φ.	(1), Symmetry
(3)	It's knowable to me that I know that φ.	(1), WEAK KK
(4)	It's knowable to you that I know that φ.	(3), Symmetry
(5)	It's knowable to you that you know that φ.	(2), WEAK KK
(6)	It's knowable to me that you know that φ.	(5), Symmetry

[13] On why common knowledge is required and simply positing second-order mutual knowledge, or third-order mutual knowledge, and so on doesn't always suffice in explaining successes in coordination, see Greco (2014b).

[14] See Greco (2015, 2016).

So, under conditions of symmetry, WEAK KK makes second-order mutual knowledge possible (even if not always attained). It is not difficult to see how the reasoning can be generalised to show how WEAK KK makes mutual knowledge of any order and thus common knowledge possible.[15]

Again, bracketing off whatever misgivings one might have about the argument, the important thing to note for present purposes is that the argument also seems to provide some motivation for WEAK K¬K. For, consider a variation of the previous example:

> As before, members of a task force stationed at different terminals have to coordinate to deactivate a bomb. But now, one of four things can happen:
>
> (i) Everyone enters the correct code, which deactivates the bomb;
>
> (ii) Everyone enters the same *incorrect* code, in which case the terminals shut down briefly before deactivation can be attempted again;
>
> (iii) Different codes are entered or some but not all enter a code, in which case the bomb immediately detonates;
>
> (iv) No codes are entered, in which case nothing happens.
>
> Members of the task force receive two messages from headquarters in quick succession. The first, sent to the entire task force through the open channel, reads: 'the code is 0902'. But the message that immediately follows reads: 'Disregard the previous message. The source was unreliable'. Compare:
>
> **GROUP*.** The second message, like the first, is also sent to the entire task force as a group through the open channel. Reasonably expecting the others to have seen the second message retracting the first, members of the task force do nothing and await further instructions.
>
> **PRIVATE*.** Through a lack of foresight again, the second message, unlike the first, is sent as private messages to each member of the task force separately. Reasonably assuming that the retraction was mistakenly sent only to them and no one else, each member expects the others to falsely believe the code provided to be correct. They know that the bomb will go off if only some of them enter a code. So, they all enter the same incorrect code, shutting down the terminals briefly.

[15] For an argument against common knowledge, see Lederman (2018).

In this case, what explains the difference is that only in the first case is there common knowledge about the mutual *lack* of knowledge about what the code is. Not only does no one know the code but everyone knows that no one knows the code, everyone knows that everyone knows that no one knows the code, and so on.

But the parallel explanation of the possibility of common knowledge of a mutual lack of knowledge requires WEAK K$\neg$K for its first step. For, assuming in the simple case that our epistemic situation is (knowably) symmetrical and I don't know that φ:

(1)	I don't know that φ.	assumption
(2)	You don't know that φ.	(1), Symmetry
(3)	It's knowable to me that I don't know that φ.	(1), WEAK K$\neg$K
(4)	It's knowable to you that I don't know that φ.	(3), Symmetry
(5)	It's knowable to you that you don't know that φ.	(2), WEAK K$\neg$K
(6)	It's knowable to me that you don't know that φ.	(5), Symmetry

So, WEAK K$\neg$K accounts for how mutual knowledge of the mutual lack of knowledge (everyone knows that no one knows) is possible, after which WEAK KK further explains how common knowledge of mutual lack of knowledge is possible. So, as with the Moorean argument, the argument for WEAK KK provides at least some motivation for also accepting WEAK K$\neg$K.

Ultimately though, for reasons soon to be made clear, WEAK KK and WEAK K$\neg$K cannot both be accepted. So, with both the Moorean and common knowledge arguments, either the arguments for WEAK KK or the alleged parity between those arguments and the analogous arguments for WEAK K$\neg$K must fail. Whatever the case, the key point of this section is that there is initial motivation for accepting WEAK KK and WEAK K$\neg$K separately and also jointly. This suffices to motivate interest in the result to which we now turn.

3 The Collapse Result

The central observation of this paper is that WEAK KK and WEAK K$\neg$K jointly entail KK and K$\neg$K. More precisely, our base logic is the smallest normal modal logic in which knowledge is factive ($K\varphi \rightarrow \varphi$, or equivalently, $\varphi \rightarrow M\varphi$).[16] The result which we now prove is that any normal extension of the base logic that contains WEAK KK and WEAK K$\neg$K as theorems must also contain KK and K$\neg$K as

[16] A normal modal logic contains the theorems of the propositional calculus, the distribution axioms for K and $\Box$ ($K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$ and $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$) and is closed under the Necessitation Rules for K and $\Box$ (if $\vdash \varphi$ then $\vdash K\varphi$ and $\vdash \Box\varphi$), Modus Ponens, and uniform substitution. As we note later, the assumption of normality is merely for simplicity—weaker assumptions also suffice.

theorems.[17]

First, note that normality allows strings of K and its dual M to be distributed over conjunctions:

(K-DISTRIBUTION) $O(\varphi \wedge \psi) \rightarrow (O\varphi \wedge O\psi)$, where O is a string of K's and M's.

It follows that if WEAK KK and WEAK K¬K are theorems, then so is:

$$\varphi \rightarrow KKM\varphi. \qquad (\star)$$

Suppose not for a contradiction:

(1)	$\varphi \wedge \neg KKM\varphi$	assumption
(2)	$M(\varphi \wedge \neg KKM\varphi)$	(1), K-FACTIVITY
(3)	$\Diamond KM(\varphi \wedge \neg KKM\varphi)$	(2), WEAK K¬K
(4)	$\Diamond\Diamond\Diamond KKKM(\varphi \wedge \neg KKM\varphi)$	(3), WEAK KK
(5)	$\Diamond\Diamond\Diamond(KKKM\varphi \wedge KKKM\neg KKM\varphi)$	(4), K-DISTRIBUTION
(6)	$\Diamond\Diamond\Diamond(KKKM\varphi \wedge KKK\neg KKKM\varphi)$	(5), ¬-ELIMINATION
(7)	$\Diamond\Diamond\Diamond(KKKM\varphi \wedge \neg KKKM\varphi)$	(6), K-FACTIVITY

Since (7) is a contradiction, $(\star)$ is a theorem. KK is then immediate:

(1)	$\vdash \varphi \rightarrow KM\varphi$	$(\star)$, K-FACTIVITY
(2)	$\vdash MK\varphi \rightarrow \varphi$	(1), DUALITY[18]
(3)	$\vdash KKMK\varphi \rightarrow KK\varphi$	(2), NORMALITY
(4)	$\vdash K\varphi \rightarrow KKMK\varphi$	$(\star)$, UNIFORM SUBSTITUTION
(5)	$\vdash K\varphi \rightarrow KK\varphi$	(3), (4)

K¬K then follows from the well-known fact that K¬K (the epistemic 5-axiom) follows from (1) (the epistemic B-axiom) and KK (the epistemic 4-axiom).[19] This completes the proof that KK and K¬K are theorems if WEAK KK and WEAK K¬K are.[20]

Initially, the result seems surprising. WEAK KK and WEAK K¬K seem ex-

[17]Going forward, the base logic will always be assumed in the background and any proposed extension of it will be assumed to be normal. Thus, for instance, claims like 'WEAK KK is strictly weaker than KK' should be understood as the claim that the smallest normal extension of the base logic containing WEAK KK does not contain KK.

[18]Where O is a (possibly empty) string of K's and M's, let $\hat{O}$ be the dual string where any occurrence of K or M is replaced by its dual (e.g. if $O = KMK$, then $\hat{O} = MKM$). According to DUALITY, $O_1\varphi \rightarrow O_2\psi$ is a theorem just in case $\hat{O}_2\psi \rightarrow \hat{O}_1\varphi$ is. DUALITY holds in all normal modal logics.

[19]For instance, see Chellas (1980, 137).

[20]The background assumptions required are fairly uncontroversial. The factivity of knowledge is beyond dispute and the assumption of normality, while contestable, can be weakened significantly, as shown in a technical companion to this paper. The proof is too involved to reproduce here but interested readers are referred to San (2019)—especially §3 and Theorem 11.

tremely weak—so weak as to appear palatable to even the most hardline of KK and K¬K skeptics. This is so especially on certain readings of $\Diamond$ (consider, for instance, 'it is metaphysically possible that' or 'it is permissible that'). But the surprise diminishes upon reflection on a closely related result. The Church–Fitch Theorem (better known as 'Fitch's Paradox' or the 'Paradox of Knowability') shows that the Knowability Principle:

(KP) If φ, then possibly one knows that φ ($\varphi \to \Diamond K\varphi$)

entails the Omniscience Principle:

(OP) If φ, then one knows that φ ($\varphi \to K\varphi$).[21]

The proof is simple:

(1)	$\Diamond K(\varphi \wedge \neg K\varphi)$	assumption
(2)	$\Diamond(K\varphi \wedge K\neg K\varphi)$	(1), K-DISTRIBUTION
(3)	$\Diamond(K\varphi \wedge \neg K\varphi)$	(2), K-FACTIVITY
(4)	$\vdash \neg\Diamond K(\varphi \wedge \neg K\varphi)$	(1), (3), *reductio*
(5)	$\vdash (\varphi \wedge \neg K\varphi) \to \Diamond K(\varphi \wedge \neg K\varphi)$	KP, UNIFORM SUBSTITUTION
(6)	$\vdash \neg(\varphi \wedge \neg K\varphi)$	(4), (5)
(7)	$\vdash \varphi \to K\varphi$	(6)

KP is thus stronger than it initially seems—weak assumptions suffice for it to collapse into OP.[22] In light of this, the collapse of WEAK KK and WEAK K¬K into KK and K¬K is not so unprecedented. After all, WEAK KK and WEAK K¬K are simply restrictions of KP to certain propositions. While KP says that all truths are knowable, WEAK KK says that all truths *about what is known* are knowable and WEAK K¬K says that all truths *about what is not known* are knowable. Similarly, KK and K¬K are also restrictions of OP to truths about what is known and truths about what is not known, respectively. Like the Church–Fitch Theorem, our collapse result shows that weak assumptions suffice to, in some sense, erase the logical distinction between possible knowledge and actual knowledge.

The collapse result has important philosophical upshots. Call any principle that entails WEAK KK a *positive introspection* principle and any principle that entails WEAK K¬K a *negative introspection* principle.[23] The collapse result above

[21] This result is often attributed to Fitch (1963) but was first noted by Church (2009) in an anonymous referee report.

[22] In fact, the assumptions required for the collapse is much weaker than the ones assumed here. In particular, while the standard derivation of the Church–Fitch Theorem appeals to the factivity of K, this assumption can be weakened significantly (see San (2020)).

[23] A popular variant of KK states that if one knows that φ, then one is *in a position to know* that one knows that φ. Whether this counts as a positive introspection principle in the sense stipulated above (i.e. whether being in a position to know entails the possibility of knowing, on some notion of possibility) is somewhat unclear. Some, like Heylen (2016, 64), are explicit about being in a position to know being a stronger notion than the possibility of knowing. For others

severely constrains the space of tenable combinations of introspection principles. On pain of accepting K¬K, at least one of WEAK KK and WEAK K¬K must be rejected. Proponents of positive introspection are thus forced to reject WEAK K¬K. Skeptics of KK, on the other hand, are faced with three possibilities: accept only WEAK KK, accept only WEAK K¬K, or accept neither. Adjudicating on these possibilities requires a better understanding of the individual strength of WEAK KK and WEAK K¬K. This is the goal of the next two sections. (We will focus on the metaphysical interpretation of possibility for concreteness, though the discussion may generalise to various other interpretations).

4 Positive Introspection

The collapse result raises some questions about WEAK KK. How strong is WEAK KK on its own? Does it, by itself, already give rise to KK?[24] If not, what do models that validate WEAK KK but not KK look like? Do standard arguments against KK also tell against WEAK KK? This section aims to shed light on these questions. I show that although WEAK KK is genuinely logically weaker than KK, it is susceptible to a variant of the influential margin-of-error argument against KK.

The margin-of-error argument against KK involves cases of imperfect discrimination.[25] Consider a tree whose height is being estimated at a distance. From that distance, one's vision is imperfect but not so radically impaired as to not be able to discriminate between trees of any height whatsoever. For concreteness, assume that one can reliably tell apart height differences of more than two inches but not any less.[26] Thus, for all $n \geq 1$, if the tree is $n-1$ inches, for all one knows, it is n inches instead ($\mathbf{n-1} \rightarrow M\mathbf{n}$, where $\mathbf{n}$ is the proposition that the tree is n inches tall).[27] Put differently, if one knows that the tree isn't n inches, then it isn't $n-1$ inches ($K\neg\mathbf{n} \rightarrow \neg\mathbf{n-1}$). Suppose that this margin-of-error principle

like (Cohen, 1999, 84), (Williamson, 2000, 95), and (Stanley, 2008, 49), this is implicit in what they say about being in a position to know. Certain arguments for the in-a-position-to-know variant of KK, like the argument from dubious assertions (see Greco (2014b, 2015)), also seem to presuppose that being in a position to know entails the possibility of knowing. However, some, like Spencer (2017), explicitly deny this entailment. No attempt to resolve this unclarity will be made in this paper—I leave that burden up to each author to specify what they mean by 'being in a position to know'.

[24]Clearly, WEAK KK does not by itself give rise to K¬K. Intuitively, this is because WEAK KK, on natural interpretations of $\Diamond$, is at most as strong as KK and KK does not imply K¬K. A more formal proof, which basically makes this intuitive thought rigorous, would make use of the conservative extension result of Thomason (1980) (see San (2019), Theorem 20).

[25]See Williamson (2000, ch. 5).

[26]We concern ourselves with cases in which the tree is in fact whatever height it appears to be. For cases in which appearance and reality diverge, see Williamson (2013a).

[27]The margin-of-error principle that $\mathbf{n+1} \rightarrow M\mathbf{n}$ is, of course, equally possible. For simplicity, we focus on the margin-of-error principle above.

is known:

$$K(K\neg\mathbf{n} \rightarrow \neg\mathbf{n}-\mathbf{1}), \quad \text{for all } n \geq 1. \tag{$*$}$$

Now, suppose the tree is 100 inches tall. Given the setup, one knows that the tree is not 102 inches tall. But then contradiction ensues given KK:

(1)	$K\neg\mathbf{102}$	assumption
(2)	$KK\neg\mathbf{102}$	(1), KK
(3)	$K(K\neg\mathbf{102} \rightarrow \neg\mathbf{101})$	$(*)$, $n = 102$
(4)	$KK\neg\mathbf{102} \rightarrow K\neg\mathbf{101}$	(3), NORMALITY
(5)	$K\neg\mathbf{101}$	(2), (4)
(6)	$KK\neg\mathbf{101}$	(5), KK
(7)	$K(K\neg\mathbf{101} \rightarrow \neg\mathbf{100})$	$(*)$, $n = 101$
(8)	$KK\neg\mathbf{101} \rightarrow K\neg\mathbf{100}$	(7), NORMALITY
(9)	$K\neg\mathbf{100}$	(6), (8)
(10)	$\neg\mathbf{100}$	(9), K-FACTIVITY

Thus, KK is incompatible with knowledge of the margin-of-error principle.

Models of modal logic nicely illustrate this incompatibility. A model $\mathfrak{M} = \langle W, R_K, R_\Box, V\rangle$ consists of a non-empty set W of 'worlds', binary accessibility relations R_K and $R_\Box$ between worlds (the epistemic and metaphysical accessibility relations), and a valuation function V specifying the worlds at which each atomic sentence is true. Whenever wR_Kv, we say that w *K-accesses* v. It is true at w that one knows that φ just in case φ is true at every world K-accessed by w. Similarly, whenever $wR_\Box v$, we say that w $\Box$*-accesses* v. And it is true at w that necessarily φ just in case φ is true at every world $\Box$-accessed by w.[28]

A model of the case just described is depicted in Figure 1 (the arrows represent R_K).[29] The case's key feature of imperfect discrimination is encoded in the

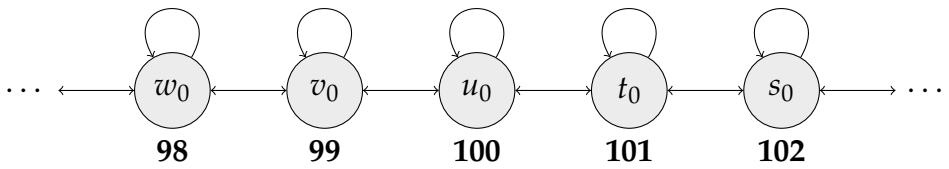

Figure 1: Margin-of-Error Model

non-transitivity of R_K.[30] For instance, 100-inch trees might be 101 inches and 101-inch trees might be 102 inches but 100-inch trees are known not to be 102 inches.

[28]For precise definitions of the familiar notions of validity in a model, truth at a world in a model, and so on, see the Appendix.

[29]For simplicity, $R_\Box$ is henceforth always omitted and assumed to be universal, i.e. every world $\Box$-accesses every other.

[30]R_K is transitive (over W) if $\forall w, v, u \in W(wR_Kv \wedge vR_Ku \rightarrow wR_Ku)$.

This is reflected in the fact that $u_0 R_K t_0$ and $t_0 R_K s_0$ but not $u_0 R_K s_0$. Failures of transitivity like this give rise to failures of KK (at u_0, $K\neg\mathbf{102} \wedge \neg KK\neg\mathbf{102}$).

Could a variant of the margin-of-error argument against KK be made against WEAK KK?[31] To be sure, the model above invalidates not only KK but also WEAK KK.[32] However, this is not yet an argument against WEAK KK. Models are often simplified in that worlds are individuated only to the extent practically necessary. In particular, models like the one above which are designed to probe purely epistemic principles like KK may justifiably leave out peripheral metaphysical possibilities that have no bearing on the tenability of KK. However, these possibilities become salient when evaluating metaphysically qualified principles like WEAK KK.

A natural first step in assessing the strength of WEAK KK relative to KK is thus to consider whether a finer individuation of worlds can help validate WEAK KK. Indeed, it can. A general way of modifying a non-transitive model to validate WEAK KK (without making the model transitive) is to append to each world a transitively-closed chain of worlds, as in Figure 2.[33] Even without yet specifying what is true at the newly added worlds, it can already be shown that any such model validates WEAK KK. For just as the factivity of knowledge is well-known to correspond to the reflexivity of R_K, KK to its transitivity, and so on, WEAK KK corresponds to the condition that:

$$\forall w \exists v (w R_\Box v \wedge \forall u t ((v R_K u \wedge u R_K t) \rightarrow w R_K t)).^{34}$$

In words: every world w $\Box$-accesses a world v such that every world K-accessible from v in two-steps is K-accessed by w. The enriched model satisfies this condition and thus validates WEAK KK.[35]

[31] Williamson takes the margin-of-error argument to be an argument against the in-a-position-to-know variant of KK, where being in a position to know p means knowing p once one 'has done what one is in a position to do to decide whether p is true' (2000, 95). However, as already noted, there is substantial variation in how the notion of being in a position to know is understood in the literature. Furthermore, the margin-of-error against the in-a-position-to-know variant of KK seems to leave certain variants of WEAK KK entirely untouched (for instance, it says nothing about the deontic variant of WEAK KK where $\Diamond$ is interpreted as 'it is permissible that'). The margin-of-error argument against WEAK KK outlined later in this section has the benefit of circumventing the unclarity of the notion of being in a position to know as well as being applicable to all variants of WEAK KK.

[32] This is not an artefact of our simplifying assumption that $R_\Box$ is universal. That no specification of $R_\Box$ will validate WEAK KK is apparent from the fact that $K(\neg\mathbf{102} \wedge \neg KKK\neg\mathbf{102})$ is true at u_0 but $\Diamond\Diamond KKK(\neg\mathbf{102} \wedge \neg KKK\neg\mathbf{102})$ is a contradiction.

[33] A subset $X \subseteq W$ of worlds is transitively-closed (relative to R_K) if R_K is transitive over X. The universal $R_\Box$ and reflexive R_K arrows in Figure 2 are omitted.

[34] See the Appendix for the proof.

[35] Every newly added world satisfies the condition, since these all belong to transitively-closed chains. And every world from the original model also satisfies it—for instance, w_0 $\Box$-accesses w_1 and every world K-accessible from w_1 in two-steps is K-accessed by w_0 because, by construction, $w_0, w_1, w_2, \ldots$ is transitively closed.

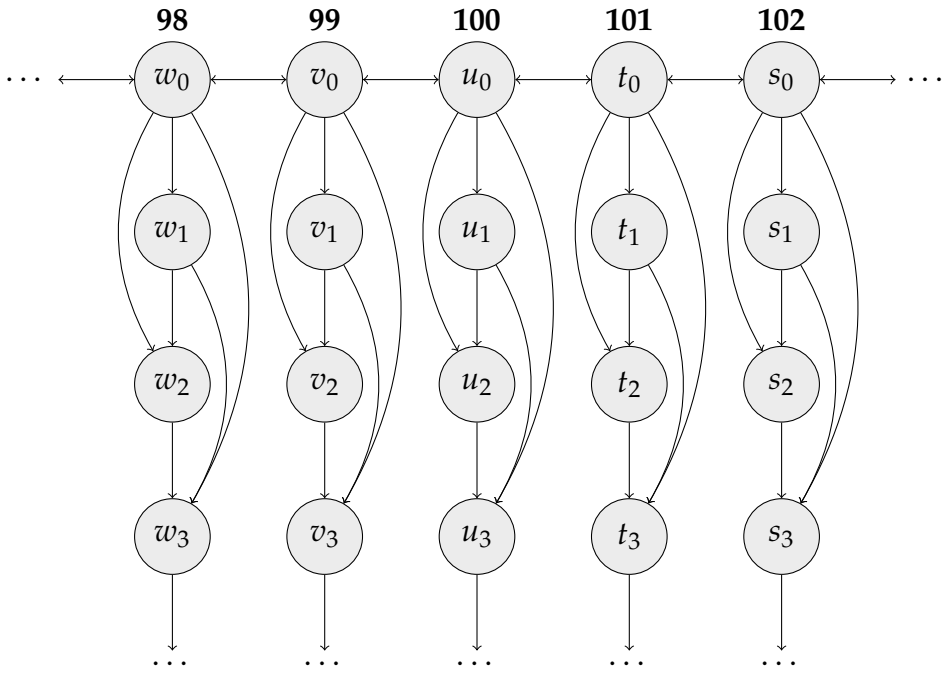

Figure 2: Enriched Margin-of-Error Model

Certain specifications of what is true at the newly added worlds however, will invalidate KK. For instance, if **98** is not true at any u_i ($i \geq 0$), then $K\neg\mathbf{98} \wedge \neg KK\neg\mathbf{98}$ will be true at u_0. Thus, there are models that validate WEAK KK while invalidating KK. This means that relative to the base logic, WEAK KK is genuinely logically weaker than KK. Thus, WEAK KK alone is insufficient for deriving the collapse result of the previous section.

The enriched models which validate WEAK KK but not KK provide some insight into how to extend the margin-of-error argument to WEAK KK. Careful inspection of such models reveals that they are, intuitively, models in which limits to the agent's discriminatory powers are epistemically contingent. That is, one's powers of discrimination, though limited, are not so for all one knows.[36] Indeed, the epistemic contingency of the limits to the agent's discriminatory powers is, in some sense, an essential feature of models that validate WEAK KK but not KK.

[36]To see why, first note that in enriched models that validate WEAK KK but not KK, there has to be an $\mathbf{n}$ that is false everywhere along some chain. For instance, for u_0 to be a witness to KK-failure of the form $K\neg\mathbf{98} \wedge \neg KK\neg\mathbf{98}$, it must be that **98** is false at every u_i ($i \geq 0$). In turn, this means that the margin-of-error principle must fail at some u_i.

For, suppose without loss of generality that $\mathbf{n}$ is false everywhere along the chain $(u_i)_{i\geq 1}$. Then, $\mathbf{n}-\mathbf{1} \rightarrow M\mathbf{n}$ will be false at any u_i where $\mathbf{n}-\mathbf{1}$ is true. If $\mathbf{n}-\mathbf{1}$ also happens to be false at every u_i, then $\mathbf{n}-\mathbf{2} \rightarrow M\mathbf{n}-\mathbf{1}$ would be false at any u_i where $\mathbf{n}-\mathbf{2}$ is true. And so on. Equally, $\mathbf{n}+\mathbf{1} \rightarrow M\mathbf{n}$ would be false at any u_i where $\mathbf{n}+\mathbf{1}$ is true. And so on. Whatever the case, *some* instance of the margin-of-error principle must fail somewhere along the chain. And since u_0 K-accesses every u_i, there must be some instance of the margin-of-error principle such that at u_0, it fails for all one knows (i.e. for some n, $M(\mathbf{n}-\mathbf{1} \wedge K\neg\mathbf{n})$). So, at u_0, one's powers of discrimination, though limited, are not so for all one knows.

For, just as KK is inconsistent with knowledge of the margin-of-error principle, so too WEAK KK can be shown to be inconsistent with *higher-order* knowledge of the margin-of-error principle.

Let nth-order knowledge (K^n) be knowing that... knowing that one knows (n times). Assume that the margin-of-error principle is not just known but known up to an arbitrary order:

$$K^{n+1}(K\neg\mathbf{i}-\mathbf{1}\rightarrow\neg\mathbf{i}),\quad \text{for any } n\geq 0, i\geq 1.^{37}$$

Distributing the inner K^n over the conditional yields:

$$K(K^{n+1}\neg\mathbf{i}-\mathbf{1}\rightarrow K^n\neg\mathbf{i}).$$

Then, contraposing the embedded conditional:

$$K(M^n\mathbf{i}\rightarrow M^{n+1}\mathbf{i}-\mathbf{1}).$$

Instances of this include:

$K(M^m\mathbf{j}+\mathbf{k}\rightarrow M^{m+1}\mathbf{j}+\mathbf{k}-\mathbf{1})$	$(i=j+k, n=m)$
$K(M^{m+1}\mathbf{j}+\mathbf{k}-\mathbf{1}\rightarrow M^{m+2}\mathbf{j}+\mathbf{k}-\mathbf{2})$	$(i=j+k-1, n=m+1)$
...	...
$K(M^{m+k-2}\mathbf{j}+\mathbf{2}\rightarrow M^{m+k-1}\mathbf{j}+\mathbf{1})$	$(i=j+2, n=m+k-2)$
$K(M^{m+k-1}\mathbf{j}+\mathbf{1}\rightarrow M^{m+k}\mathbf{j})$	$(i=j+1, n=m+k-1)$

By the closure of knowledge under known implication, it follows that:

$$K\left(M^m\mathbf{j}+\mathbf{k}\rightarrow M^{m+k}\mathbf{j}\right),\quad \text{for all } m,j,k\geq 0. \qquad (**)$$

For instance, where $m=1, j=1$, and $k=99$, $K(M\mathbf{100}\rightarrow M^{100}\mathbf{1})$. That is, one knows that lacking first-order knowledge that the tree isn't 100 inches tall means lacking 100th-order knowledge that it isn't an inch tall.

Now, given a suitable description of the situation, it is plausible that one knows both that the tree isn't an inch tall and that for all one knows, it is 100 inches tall. But this leads to a contradiction given ($**$) and WEAK KK:

(1)	$K(\neg\mathbf{1}\wedge M\mathbf{100})$	assumption
(2)	$K\left(M\mathbf{100}\rightarrow M^{100}\mathbf{1}\right)$	(**), $m=1, j=1, k=99$
(3)	$K(\neg\mathbf{1}\wedge M^{100}\mathbf{1})$	(1), (2)
(4)	$K(\neg\mathbf{1}\wedge\neg K^{100}\neg\mathbf{1})$	(3)

[37] K^n and M^n abbreviate $n\geq 0$ iterations of K and M, respectively (they are the empty string when $n=0$). We assume that the margin-of-error principle is known up to an arbitrary order for generality. As should be clear from the subsequent paragraphs, the argument against WEAK KK does not require that knowledge of the margin-of-error principle iterate indefinitely but only up to some sufficiently high level.

(5) $\Diamond^{99}K^{100}(\neg\mathbf{1} \wedge \neg K^{100}\neg\mathbf{1})$ (4), WEAK KK

(6) $\Diamond^{99}(K^{100}\neg\mathbf{1} \wedge K^{100}\neg K^{100}\neg\mathbf{1})$ (5), K-DISTRIBUTION

(7) $\Diamond^{99}(K^{100}\neg\mathbf{1} \wedge \neg K^{100}\neg\mathbf{1})$ (6), K-FACTIVITY

The upshot is that just as KK is inconsistent with knowledge of the margin-of-error principle, so too WEAK KK is inconsistent with higher-order knowledge of the margin-of-error principle.

Thus, although accepting WEAK KK while rejecting KK is logically consistent, its philosophical tenability is another matter. For those who reject KK on margin-of-error grounds, the tenability of WEAK KK rests on the plausibility of higher-order knowledge of the margin-of-error principle as opposed to mere first-order knowledge of it. For instance, it would be natural for proponents of WEAK KK who reject KK to resist the argument against WEAK KK by accepting only the *possibility* of higher-order knowledge of the margin-of-error principle.[38] A proper assessment of this and other responses goes beyond an exploratory paper of the present kind and is best left as a future undertaking. For now, we must content ourselves with having clarified some of the issues upon which any difference in stance on KK and WEAK KK depends.[39]

5 Negative Introspection

The questions that arise for WEAK KK could also be asked of WEAK K¬K. How strong is WEAK K¬K on its own? Does it, by itself, already give rise to K¬K or KK? If not, what do models that validate WEAK K¬K but not K¬K look like? Do standard arguments against K¬K also tell against WEAK K¬K? This section aims to shed light on these questions. I show that although WEAK K¬K is genuinely logically weaker than K¬K and KK, it is susceptible to a variant of the influential anti-skeptical argument against K¬K.

A distinctive characteristic of being in a skeptical scenario is that one does not know that one is in a skeptical scenario. As Williamson puts it, 'part of the badness of the bad case is that one cannot know just how bad one's case is' (2000, 165). That is:

$$\mathbf{b} \rightarrow \neg K\mathbf{b}, \tag{†}$$

where $\mathbf{b}$ is the proposition that one is a brain-in-a-vat (or any other skeptical claim of one's choice). The disagreement between skeptics and anti-skeptics lies in whether ignorance of whether one is in a brain-in-a-vat extends to the non-envatted. Anti-skeptics hold that under suitable conditions, the non-envatted know that they are not brains-in-vats. And presumably, anti-skeptics take that

[38] Thanks to an anonymous referee for this observation.

[39] See Liu (2020) on how other arguments against KK can also be extended to WEAK KK.

not just to be true but *known* to be true:

$$K(\neg \mathbf{b} \rightarrow K\neg \mathbf{b}). \qquad (\dagger\dagger)$$

The anti-skeptical argument against K¬K stems from the observation that $(\dagger)$ and $(\dagger\dagger)$ imply that K¬K fails in skeptical scenarios:[40]

(1)	$K(\neg \mathbf{b} \rightarrow K\neg \mathbf{b})$	$(\dagger\dagger)$
(2)	$K(\neg K\neg \mathbf{b} \rightarrow \mathbf{b})$	(1), NORMALITY
(3)	$K\neg K\neg \mathbf{b} \rightarrow K\mathbf{b}$	(2), NORMALITY
(4)	$\mathbf{b} \rightarrow \neg K\mathbf{b}$	$(\dagger)$
(5)	$\mathbf{b} \rightarrow \neg K\neg K\neg \mathbf{b}$	(3), (4)
(6)	$\mathbf{b} \rightarrow \neg K\neg \mathbf{b}$	K-FACTIVITY

Thus, according to (5) and (6), an agent in the bad case doesn't know that she isn't in the bad case but *contra* K¬K, also doesn't know that she doesn't know that.

This violation of K¬K can again be nicely illustrated with a model (see Figure 3). The non-symmetry of R_K encodes an anti-skeptical stance.[41] The fact that

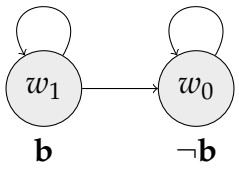

Figure 3: Anti-Skeptical Model

the envatted don't know that they are brains-in-vats is reflected in the fact that $w_1 R_K w_0$. On the other hand, the anti-skeptical claim that the non-envatted know that they are not brains-in-vats is captured by the fact that it is not the case that $w_0 R_K w_1$. This failure of symmetry gives rise to a failure of K¬K (at w_1, $\neg K\neg \mathbf{b} \wedge \neg K\neg K\neg \mathbf{b}$).[42]

Although this model invalidates WEAK K¬K in addition to K¬K, this should not yet be taken to be an argument against WEAK K¬K for a familiar reason.[43] The worlds may simply not be individuated finely enough. A proper assessment of WEAK K¬K requires first considering whether the anti-skeptical model can be modified to validate WEAK K¬K without validating K¬K. One way (though not the only way) is to add an infinite chain of worlds as in Figure 4.[44] Note that

[40] For this presentation of the argument, I am greatly indebted to an anonymous referee.

[41] R_K is symmetric (over W) if $\forall w, v \in W(wR_K v \rightarrow vR_K w)$.

[42] See Williamson (2000, ch. 8).

[43] The model invalidates WEAK K¬K because it is true at w_1 that $\neg K\neg \mathbf{b}$ but *contra* WEAK K¬K,

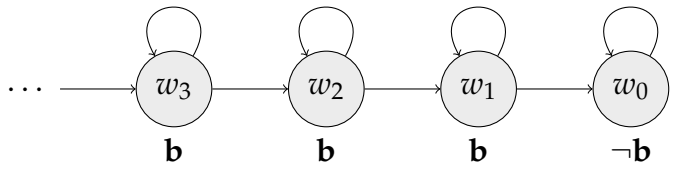

Figure 4: Enriched Anti-Skeptical Model

WEAK K¬K corresponds to the condition that:

$$\forall wv(wR_Kv \rightarrow \exists u(wR_\Box u \wedge \forall t(uR_Kt \rightarrow tR_Kv))).^{45}$$

In words: if w K-accesses v, then it $\Box$-accesses some world u such that v is K-accessed by every world that u K-accesses. The enriched model satisfies this condition and thus validates WEAK K¬K.[46]

While the model validates WEAK K¬K, it invalidates K¬K, since $\neg K\neg\mathbf{b} \wedge \neg K\neg K\neg\mathbf{b}$ is still true at w_1. And it also invalidates KK, since $K\mathbf{b} \wedge \neg KK\mathbf{b}$ is true at w_2. This shows that WEAK K¬K is genuinely logically weaker than both KK and K¬K. Thus, responsibility for the collapse result cannot be attributed solely to either WEAK KK or WEAK K¬K. KK and K¬K are the product of a genuine synergy between their 'weak' counterparts.

Although the enriched model shows that WEAK K¬K is logically weaker than K¬K, it also provides a clue as to how the anti-skeptical argument might be extended to WEAK K¬K. Intuitively, the newly added worlds are metaphysical possibilities in which the envatted know of their predicament. Thus, in the enriched model, w_1 is a world in which although the envatted don't know that they are brains-in-vats, it is nevertheless *possible* that they do.

Indeed, this is an essential feature of models of WEAK K¬K that invalidate K¬K. Given (††), just as K¬K is inconsistent with (†), so too WEAK K¬K is inconsistent with a modal strengthening of (†). According to the modal strengthening, it is *impossible* for the envatted to know of their predicament:

$$\mathbf{b} \rightarrow \neg\Diamond K\mathbf{b}. \qquad (\Box\dagger)$$

Given (††), ($\Box$†) entails the failure of WEAK K¬K for brains-in-vats:

$\Diamond K\neg K\neg\mathbf{b}$ is not no matter what the $\Box$-accessibility relations are among w_1 and w_2.

[44] $R_\Box$ arrows are, as always, omitted and assumed to be universal. For another way of modifying the model to validate WEAK K¬K while invalidating K¬K, see Theorem 22 of San (2019).

[45] See the Appendix for the proof.

[46] In the enriched model, by construction, every world K-accesses only itself and (except for w_0) the world immediately to its right. That is, the only accessibility relations are $w_nR_Kw_n$ (for all $n \geq 0$) and $w_nR_Kw_{n-1}$ (for all $n \geq 1$). Thus, the condition corresponding to WEAK K¬K is satisfied because $w_nR_\Box w_{n+1} \wedge \forall t(w_{n+1}R_Kt \rightarrow tR_Kw_n)$ (for all $n \geq 0$) and $w_nR_\Box w_n \wedge \forall t(w_nR_Kt \rightarrow tR_Kw_{n-1})$ (for all $n \geq 1$).

(1)	$K(\neg \mathbf{b} \rightarrow K\neg \mathbf{b})$	$(\dagger\dagger)$
(2)	$K(\neg K\neg \mathbf{b} \rightarrow \mathbf{b})$	(1), NORMALITY
(3)	$\Diamond K\neg K\neg \mathbf{b} \rightarrow \Diamond K\mathbf{b}$	(2), NORMALITY
(4)	$\mathbf{b} \rightarrow \neg\Diamond K\mathbf{b}$	$(\Box\dagger)$
(5)	$\mathbf{b} \rightarrow \neg\Diamond K\neg K\neg \mathbf{b}$	(3), (4)
(6)	$\mathbf{b} \rightarrow \neg K\neg \mathbf{b}$	K-FACTIVITY

Thus, according to (5) and (6), an agent in the bad case doesn't know that she isn't in the bad case but *contra* K$\neg$K, it's also impossible for her to know that she doesn't know that.

Of course, for proponents of positive introspection, any argument against K$\neg$K is automatically an argument against WEAK K$\neg$K given the collapse result. So, this variant of the argument explicitly targeting WEAK K$\neg$K wouldn't be of much interest to them. For those who reject positive introspection, however, their decision on whether to reject WEAK K$\neg$K in addition to K$\neg$K might depend on how plausible $(\Box\dagger)$ is relative to $(\dagger)$. As before, a proper assessment of this is best left for a future investigation but some cursory remarks might bring out why this is not a straightforward matter.[47]

The reason that $(\Box\dagger)$ fails at w_1 is because $w_2, w_3, \ldots$ are metaphysically accessible $\mathbf{b}$-worlds in which $\mathbf{b}$ is known. This model is not at all unrealistic if $\mathbf{b}$ is simply the proposition that one is a brain-in-a-vat. After all, there is no inherent contradiction in knowing that one's brain is suspended in a vat of liquid. For instance, there are metaphysical possibilities in which one is a brain-in-a-vat with perfectly reliable faculties. One might clarify that 'being a brain-in-a-vat' is really elliptical for something like 'being a brain-in-a-vat with unreliable sensory faculties'. Still, there is no inherent contradiction in knowledge of this. The metaphysical possibility of knowing this via other means (say, divine revelation) is not ruled out. In general, whatever further qualifications would still leave open metaphysical possibilities in which the skeptical effects of the situation is negated by some other source of knowledge.

Ruling out models that validate WEAK K$\neg$K thus requires being able to identify a skeptical proposition $\mathbf{b}$ which is *impossible* to know.[48] A trick one might attempt is to simply let $\mathbf{b}$ be the proposition that one is in a bad case, where a

[47] The discussion in this section focused primarily on WEAK K$\neg$K on the metaphysical interpretation of $\Diamond$. The plausibility of $(\Box\dagger)$ and the argument against WEAK K$\neg$K might vary depending on the interpretation under consideration. For instance, under the epistemic interpretation of $\Diamond$, $(\Box\dagger)$ amounts to $\mathbf{b} \rightarrow K\neg K\mathbf{b}$. Examples in which this is true are easily found. For instance, in cases where the envatted don't believe that they are brains-in-vats and know that they don't believe that, they will also know that they don't know that they are brains-in-vats. Such cases provide straightforward counterexamples to EPIST K$\neg$K, as helpfully pointed out by an anonymous referee.

[48] There are true Moorean conjunctions of the form '$\varphi \wedge \neg K\varphi$' that are impossible to know. Thus, if we let $\mathbf{b}$ be of the form '$\varphi \wedge \neg K\varphi$', then $(\Box\dagger)$ is a theorem. But the anti-skeptical argument also requires $(\dagger\dagger)$ and it is questionable whether there is a choice of $\mathbf{b}$ of that form that makes $(\dagger\dagger)$ plausible.

bad case is stipulated to just be a case in which one doesn't know that one is in a bad case. This stipulation would make $\neg K\mathbf{b}$ true by definition and thus true necessarily. But this trick is highly suspect, for reasons that are familiar from the liar paradox and the knower paradox.[49]

6 Some Implications

Besides shedding light on the relationship between KK, K¬K, and their 'weak' counterparts, the collapse result also has potentially interesting implications for a range of issues—from the debate on positive introspection to deeper epistemological issues to do with the structure and limits of knowability. This section briefly notes some of these implications.

An immediate upshot of the collapse result is that on pain of accepting K¬K, proponents of positive introspection must reject WEAK K¬K. This places some burden on the proponents of the arguments from Moorean assertions and common knowledge to explain why, *contra* §2, those arguments do not carry over to WEAK K¬K. More generally, a moral of the collapse result is that arguments in favour of WEAK KK had better not also be arguments for WEAK K¬K.

The collapse result also has interesting implications for the structure and limits of knowability. Consider the set containing all and only everything one knows. A set is *semi-decidable* if there is a semi-decision procedure for it (an algorithm that, given an input, outputs 'positive' if the input is a member of the set and outputs 'negative' or does not halt otherwise). A set is *decidable* if both it and its complement are semi-decidable. Now, suppose there is some appropriate notion of knowability that corresponds to the existence of a semi-decision procedure. That is, it is possible to know that φ is in a set just in case there exists a semi-decision procedure for the set that outputs 'positive' for φ. Then, the collapse result entails that either there does not exist a semi-decision procedure for the set containing everything known or there does not exist one for the complement set containing everything unknown, for at least one of WEAK KK and WEAK K¬K must fail. The upshot is that the set containing everything known is not decidable. At most, it is semi-decidable.

Put differently, consider an algorithm designed to test, for any proposition, whether one knows it. Such an algorithm, if it always outputs a result, must either have false positives (it says that one knows something that one does not) or false negatives (it says that one does not know something that one does). An algorithm that is free from both would have to be such that what we learn from it cannot always culminate in knowledge. Otherwise, the existence of such an algorithm would make it possible to know of anything known that it is known and of anything unknown that it is unknown. In that case, by the collapse result, KK and K¬K would follow.

[49] See Kaplan & Montague (1960) and, for a more recent discussion, Jerzak (2019).

Introspection might be construed as an algorithm that checks whether one knows a given proposition. Consider a toy model of the mind on which anything one comes to know is stored in a mental repository. Introspecting on whether one knows φ can then be thought of as the process of mechanistically scanning the contents of the repository to see if it contains φ. What we have shown is that there must be limits to such a process. Insofar as what we learn from introspection can always culminate in knowledge, introspection cannot be completely infallible if it always delivers a result. Either it sometimes falsely proclaims something to be known or falsely proclaims something to be unknown.

This is not the mundane observation concerning the empirical limitations of human introspection. Rather, logic itself imposes restrictions on the reach of introspection. Even for creatures whose powers of introspection far exceeds ours, if it is always possible for them to know what they learn through introspection, introspection must still yield false positives or false negatives, as long as it always delivers a result and KK or K$\neg$K fails. To endow a creature with powers of introspection that are any greater is tantamount to eliminating all failures of KK and K$\neg$K. It would be nothing short of granting them omniscience about when knowledge is present or absent.[50]

7 Conclusion

To conclude on a more forward-looking note, while the literature on KK and K$\neg$K is rich, related principles governing knowledge and knowability have been relatively underexplored. While the question of whether knowing implies knowing that one knows is of interest in its own right, KK owes the place it occupies at the forefront of contemporary epistemology to its promise in helping to adjudicate on the internalist-externalist debate and on deeper issues to do with the nature of knowledge. This paper, I hope, demonstrates the equal fruitfulness of a thorough investigation into related principles governing knowledge and knowability. The consideration of principles like WEAK KK and WEAK K$\neg$K not only provides clarification of the lay of the epistemological land but also promises to shed light on the debate on positive introspection as well as on foundational epistemological issues concerning the structure and limits of knowability.[51]

[50] These logical constraints on the limits of knowability are reminiscent of the constraints on the limits of formal provability (Gödel (1931)).

[51] For helpful discussions and comments, I am grateful to Sam Carter, Simon Goldstein, Jeremy Goodman, Zach Goodsell, John Hawthorne, Matt Hewson, Ben Holguín, James Kirkpatrick, Daniel Kodsi, Sebastian Liu, Will Nalls, Laurie Paul, Richard Roth, Jeff Russell, Joe Salerno, Bernhard Salow, Gabriel Uzquiano, Tim Williamson, and Joe Zhou. Detailed comments from anonymous referees have greatly improved this paper. I also owe thanks to the audiences at Oxford's Ockham Society, USC's Speculative Society, the Pitt-CMU Philosophy Graduate Conference, and the NYU-Columbia Graduate Conference in Philosophy. Research on this paper was supported by the Ertegun Graduate Scholarship and the USC Dana & Dornsife Graduate School Fellowship.

Appendix

We prove the frame conditions that correspond to WEAK KK and WEAK K$\neg$K, respectively. A frame $\mathfrak{F}$ is an ordered-tuple $\langle W, R_\Box, R_K\rangle$, where the domain W is a non-empty set and $R_\Box \subseteq (W \times W)$ and $R_K \subseteq (W \times W)$ are binary relations on W. A model $\mathfrak{M}$ is a pair $\langle \mathfrak{F}, V\rangle$, where the valuation function V maps each propositional letter to a subset of W. If $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$, we say that $\mathfrak{M}$ is based on $\mathfrak{F}$. Truth at a world in a model is defined as usual:

$\mathfrak{M}, w \Vdash p$	iff	$w \in V(p)$, for every propositional letter p;
$\mathfrak{M}, w \Vdash \neg\varphi$	iff	not $\mathfrak{M}, w \Vdash \varphi$;
$\mathfrak{M}, w \Vdash (\varphi \wedge \psi)$	iff	$\mathfrak{M}, w \Vdash \varphi$ and $\mathfrak{M}, w \Vdash \psi$;
$\mathfrak{M}, w \Vdash \Box\varphi$	iff	for every $v \in W$ such that $\langle w, v\rangle \in R_\Box$, $\mathfrak{M}, v \Vdash \varphi$.
$\mathfrak{M}, w \Vdash K\varphi$	iff	for every $v \in W$ such that $\langle w, v\rangle \in R_K$, $\mathfrak{M}, v \Vdash \varphi$.

φ is valid in a model $\mathfrak{M}$ iff $\mathfrak{M}, w \Vdash \varphi$ for every w in the domain. φ is valid in a frame $\mathfrak{F}$ ($\mathfrak{F} \Vdash \varphi$) iff it is valid in every model based on $\mathfrak{F}$.

Proposition 1. $\mathfrak{F} \Vdash K\varphi \rightarrow \Diamond KK\varphi$ iff $\mathfrak{F}$ satisfies $\forall w \exists v(wR_\Box v \wedge \forall ut((vR_K u \wedge uR_K t) \rightarrow wR_K t))$.

Proof. ($\Rightarrow$) We prove the contrapositive. Assume $\exists w \forall v(wR_\Box v \rightarrow \exists ut(vR_K u \wedge uR_K t \wedge \neg wR_K t))$. Let w be a witness to the existential and V a valuation function on which p is true at all and only worlds K-accessed by w. Then, $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ is a countermodel to WEAK KK. In particular, $\mathfrak{M}, w \Vdash Kp \wedge \neg\Diamond KKp$. Thus, $\mathfrak{F} \nVdash K\varphi \rightarrow \Diamond KK\varphi$.

($\Leftarrow$) Let $\mathfrak{F}$ satisfy the frame condition on the right. Let $\mathfrak{M}$ be a model based on $\mathfrak{F}$ and w a world in the frame. If w K-accesses a $\neg\varphi$-world, then trivially $\mathfrak{M}, w \Vdash K\varphi \rightarrow \Diamond KK\varphi$. Thus, assume that every world K-accessed by w is a φ-world. By the frame condition, w $\Box$-accesses a v such that $\forall ut((vR_K u \wedge uR_K t) \rightarrow wR_K t))$. That is, every world that v K-accesses in two steps is K-accessed by w. Since every world K-accessed by w is a φ-world by assumption, $\mathfrak{M}, v \Vdash KK\varphi$. Thus, $\mathfrak{M}, w \Vdash K\varphi \rightarrow \Diamond KK\varphi$. Since $\mathfrak{M}$ and w were arbitrary, $\mathfrak{F} \Vdash K\varphi \rightarrow \Diamond KK\varphi$. $\Box$

Proposition 2. $\mathfrak{F} \Vdash \neg K\varphi \rightarrow \Diamond K\neg K\varphi$ iff $\mathfrak{F}$ satisfies $\forall wv(wR_K v \rightarrow \exists u(wR_\Box u \wedge \forall t(uR_K t \rightarrow tR_K v)))$.

Proof. ($\Rightarrow$) We prove the contrapositive. Assume $\exists wv(wR_K v \wedge \forall u(wR_\Box u \rightarrow \exists t(uR_K t \wedge \neg tR_K v)))$ holds for $\mathfrak{F}$. Let w, v be witnesses to the existential and V a valuation function on which $\neg p$ is true only at v. Then, $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ is a countermodel to WEAK K$\neg$K. In particular, $\mathfrak{M}, w \Vdash \neg Kp \wedge \neg\Diamond K\neg Kp$. Thus, $\mathfrak{F} \nVdash \neg K\varphi \rightarrow \Diamond K\neg K\varphi$.

($\Leftarrow$) Let $\mathfrak{F}$ satisfy the frame condition on the right. Let $\mathfrak{M}$ be a model based on $\mathfrak{F}$ and w a world in the frame. If w does not K-access any $\neg\varphi$-world, then trivially $\mathfrak{M}, w \Vdash \neg K\varphi \rightarrow \Diamond K\neg K\varphi$. Thus, assume that w K-accesses some $\neg\varphi$-world v. By the frame condition, w $\Box$-accesses a u such that $\forall t(uR_K t \rightarrow tR_K v)$.

Thus, $\mathfrak{M}, u \Vdash K\neg K\varphi$ and $\mathfrak{M}, w \Vdash \neg K\varphi \rightarrow \Diamond K\neg K\varphi$. Since $\mathfrak{M}$ and w were arbitrary, $\mathfrak{F} \Vdash \neg K\varphi \rightarrow \Diamond K\neg K\varphi$. □

References

Chellas, Brian F. 1980. *Modal Logic: An Introduction*. Cambridge University Press.

Church, Alonzo. 2009. Referee Reports on Fitch's "A Definition of Value". *Pages 13–20 of:* Salerno, Joe (ed), *New Essays on the Knowability Paradox*. Oxford University Press.

Cohen, Stewart. 1999. Contextua, lism, Skepticism, and the Structure of Reasons. *Philosophical Perspectives*, **13**, 57–89.

Das, Nilanjan, & Salow, Bernhard. 2018. Transparency and the KK Principle. *Noûs*, **52**(1), 3–23.

Dorst, Kevin. 2019. Abominable KK Failures. *Mind*, **128**(512), 1227–1259.

Edgington, Dorothy. 1985. The Paradox of Knowability. *Mind*, **94**(376), 557–568.

Edgington, Dorothy. 2010. Possible Knowledge of Unknown Truth. *Synthese*, **173**(1), 41–52.

Fara, Michael. 2010. Knowability and the Capacity to Know. *Synthese*, **173**(1), 53–73.

Fitch, Frederic B. 1963. A Logical Analysis of Some Value Concepts. *Journal of Symbolic Logic*, **28**(2), 135–142.

Gödel, Kurt. 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik*, **38**(1), 173–198.

Goodman, Jeremy, & Salow, Bernhard. 2018. Taking a Chance on KK. *Philosophical Studies*, **175**(1), 183–196.

Greco, Daniel. 2014a. Could KK Be OK? *Journal of Philosophy*, **111**(4), 169–197.

Greco, Daniel. 2014b. Iteration and Fragmentation. *Philosophy and Phenomenological Research*, **88**(1), 656–673.

Greco, Daniel. 2015. Iteration Principles in Epistemology I: Arguments For. *Philosophy Compass*, **10**(11), 754–764.

Greco, Daniel. 2016. Safety, Explanation, Iteration. *Philosophical Issues*, **26**(1), 187–208.

Heylen, Jan. 2016. Being in a Position to Know and Closure. *Thought: A Journal of Philosophy*, **5**(1), 63–67.

Hintikka, Jaakko. 1962. *Knowledge and Belief*. Ithaca: Cornell University Press.

Holliday, Wesley H. 2018. Epistemic Logic and Epistemology. *Pages 351–369 of:* Hendricks, Sven Ove Hansson Vincent F. (ed), *Introduction to Formal Philosophy*. Springer.

Jerzak, Ethan. 2019. Non-Classical Knowledge. *Philosophy and Phenomenological Research*, **98**(1), 190–220.

Kaplan, D., & Montague, R. 1960. A Paradox Regained. *Notre Dame Journal of Formal Logic*, **1**, 79–90.

Lederman, Harvey. 2018. Uncommon Knowledge. *Mind*, **127**(508), 1069–1105.

Liu, Sebastian. 2020. (Un)knowability and knowledge iteration. *Analysis*, 02. anz072.

San, Weng Kin. 2019. Disappearing Diamonds: Fitch-Like Results in Bimodal Logic. *Journal of Philosophical Logic*, **48**(6), 1003–1016.

San, Weng Kin. 2020. Fitch's Paradox and Level-Bridging Principles. *Journal of Philosophy*, **117**(1), 5–29.

Smithies, Declan. 2012. Moore's Paradox and the Accessibility of Justification. *Philosophy and Phenomenological Research*, **85**(2), 273–300.

Spencer, Jack. 2017. Able to Do the Impossible. *Mind*, **126**(502), 466–497.

Stalnaker, Robert. 2006. On Logics of Knowledge and Belief. *Philosophical Studies*, **128**(1), 169–199.

Stanley, Jason. 2008. Knowledge and Certainty. *Philosophical Issues*, **18**(1), 35–57.

Thomason, S. K. 1980. Independent Propositional Modal Logics. *Studia Logica*, **39**(2-3), 143–144.

van Ditmarsch, Hans, van der Hoek, Wiebe, & Iliev, Petar. 2012. Everything is Knowable – How to Get to Know Whether a Proposition is True. *Theoria*, **78**(2), 93–114.

Williamson, Timothy. 1996. Knowing and Asserting. *Philosophical Review*, **105**(4), 489.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

Williamson, Timothy. 2013a. Gettier Cases in Epistemic Logic. *Inquiry: An Interdisciplinary Journal of Philosophy*, **56**(1), 1–14.

Williamson, Timothy. 2013b. Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic. *Inquiry: An Interdisciplinary Journal of Philosophy*, **56**(1), 77–96.