Fallibilism and Consequence

Adam Marushak *

Journal of Philosophy
(*draft; please cite published version*)

1 Introduction

Alex Worsnip argues in favor of what he describes as a particularly robust version of fallibilism: subjects can sometimes know things that are, for them, possibly false (in the epistemic sense of "possible").[1] My aim in this paper is to show that Worsnip's argument is inconclusive for a surprising reason: the existence of possibly false knowledge turns on how we ought to model entailment or consequence relations among sentences in natural language. Since it is an open question how we ought to think about consequence in natural language, it is an open question whether there is possibly false knowledge. I close with some reflections on the relation between possibly false knowledge and fallibilism. I argue that there is no straightforward way to use linguistic data about natural language epistemic modals to either verify or refute the fallibilist thesis.

2 Worsnip on Possibly False Knowledge

Worsnip's argument for the existence of possibly false knowledge (PFK) runs as follows. Consider his example from p. 232:

*The author is grateful for helpful comments from Davide Fassio, Jie Gao, Anil Gupta, Jasmin Ozel, Doug Patterson, James Shaw, and audiences at South China Normal University and Zhejiang University.

[1] Alex Worsnip, "Possibly False Knowledge," *Journal of Philosophy*, CXII, 5 (2015): 225–246.

A: Do you know what the capital of South Africa is?

B: Yes, I think I know the answer to your question—Pretoria. But it might be Johannesburg.

B's reply sounds felicitous. Worsnip claims that in general, sentences of the following type are felicitous:

(*) I think I know that ϕ, but it is possible/it might be that not-ϕ.

Call such sentences (*)-*sentences*.[2]

Worsnip notes that (*)-sentences pose a problem for a widely held view about the relation between knowledge and epistemic modals that he calls KPL (the Knowledge/Possibility Link):

KPL: For all contexts c, $\ulcorner$it is possible that not-$\phi\urcorner$ is true at c only if $\ulcorner$I do not know that $\phi\urcorner$ is true at c.[3]

The problem is that KPL wrongly predicts that (*)-sentences should sound infelicitous. At the very least, S's assertion of $\ulcorner$I think that $\phi\urcorner$ in c commits S to $\ulcorner$it is possible that $\phi\urcorner$ being true at c. Hence, S's assertion of (*) in c commits S to (**) being true at c:

(**) It is possible that I know that ϕ, but it is possible that not-ϕ.

But if KPL is correct, then (**) is true at c only if (***) is true at c:

[2]To my knowledge, the first discussion of (*)-sentences occurs in Clayton Littlejohn, "Concessive Knowledge Attributions and Fallibilism," *Philosophy and Phenomenological Research*, LXXXIII, 3 (2011): 603–619, at p. 612. He too finds them felicitous. However, he assumes that the attitude verb takes wide scope, whereas I take it that the natural reading—and the one Worsnip has in mind—involves the conjunction scoping above the attitude verb.

[3]What I am calling KPL is what Worsnip calls CKPL, the version of KPL designed to accommodate contextualism about knowledge (see Worsnip, "Possibly False Knowledge," *op. cit.*, p. 230). I have modified his CKPL to make it clear that strictly speaking, it is sentences in contexts, not utterances of sentences, that are evaluated as true or false (see David Kaplan, "Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and Other Indexicals," in Joseph Almog, John Perry and Howard Wettstein, eds., *Themes From Kaplan* (Oxford: Oxford University Press, 1989), pp. 481–563).

(***) It is possible that I know that ϕ, but I do not know that ϕ.

(***) is an instance of what Matthew Mandelkern calls a *Wittgenstein sentence*:[4]

$$(\Diamond\phi \wedge \neg\phi)$$

Wittgenstein sentences often sound odd. For example:

(W1) # It might be raining and it is not raining.

(W2) # I might know that ϕ, but I do not know that ϕ.

If asserting a (*)-sentence involves commitment to a Wittgenstein sentence, then (*)-sentences should sound odd as well. But they don't. So KPL is false.

Finally, if KPL is false, then it follows that there is possibly false knowledge—that is, it follows that PFK is true:

> Definition 1:
> PFK: there exists some context c such that $\ulcorner$I know that $\phi\urcorner$ is true at c, and $\ulcorner$it is possible that not-$\phi\urcorner$ is true at c.[5]

[4]Wittgenstein sentences are sentences of the form $\ulcorner(\Diamond\phi \wedge \neg\phi)\urcorner$ or $\ulcorner(\neg\phi \wedge \Diamond\phi)\urcorner$ (see Matthew Mandelkern, "Bounded Modality," *Philosophical Review*, CXXVIII, 1 (2019): 1–61). The latter is what Seth Yalcin calls an *epistemic contradiction* (see Seth Yalcin, "Epistemic Modals," *Mind*, CXVI, 464 (2007): 983–1026).

[5]This definition seems to be what Worsnip has in mind when he defines PFK as follows: "S can sometimes truly assert 'it is possible that not-P' even though S knows that P" (Worsnip, "Possibly False Knowledge," *op. cit.*, p. 226). Strictly speaking, this definition entails that PFK can be verified by examples like the following: I know how to escape a maze but I don't want to tell you, so I say "the exit might be to the left, and it might be to the right" (see Andy Egan, John Hawthorne, and Brian Weatherson, "Epistemic Modals in Context," in G. Preyer and G. Peter, eds., *Contextualism in Philosophy* (Oxford: Oxford University Press, 2005), pp. 131–170, at p. 140; cf. Worsnip, "Possibly False Knowledge," *op. cit.*, pp. 244–245). But even if I speak truly, it surely does not follow that there exists possibly false knowledge in any epistemologically interesting sense relevant to fallibilism, since "might" here describes what is possible relative to *someone else's* epistemic state. I therefore assume that "might"/"possible" in PFK and KPL cannot receive the type of exocentric reading one finds in the maze case. Cf. John Hawthorne, "Knowledge and Epistemic Necessity," *Philosophical Studies*, CLVIII, 3 (2012): 493–501 on "flat-footed" uses of epistemic modals.

Importantly, however, Worsnip denies that subjects can simply claim their possibly false knowledge by asserting a so-called concessive knowledge attribution (CKA): I know that ϕ, but it is possible that not-ϕ.[6] Worsnip defends what he calls the quantifier domain restriction (QDR) account of epistemic modals: assertions update the modal base of epistemic modals. The *modal base* is a set of propositions whose intersection forms the modal's domain of quantification.[7] Possibility modals express existential quantification over the worlds in this domain: ⌜it is possible that ϕ⌝ is true at c iff there is some ϕ-world in the intersection of the modal base propositions in c. Worsnip's idea is that assertions update the modal base by adding what is asserted to the modal base of epistemic modals used in the same context. But if assertions update the modal base in this way, then ⌜I know that ϕ, but it is possible that not-ϕ⌝ expresses a contradiction in any context where it is asserted. Asserting the first conjunct updates the modal base to include the proposition that the speaker knows that ϕ. But since knowledge is factive, it follows that the intersection of the modal base propositions includes only ϕ-worlds. Hence, the second conjunct—⌜it is possible that not-ϕ⌝—is false at c, since there is no not-ϕ-world in the modal's domain of quantification. Note that QDR does not predict that (*)-sentences are similarly contradictory, for there are many worlds in which one thinks one knows that ϕ but ϕ is false. QDR thus allows Worsnip to recognize the existence of possibly false knowledge while explaining the felicity of (*)-sentences and the infelicity of CKAs.

[6]For discussion of CKAs see Keith DeRose, "Epistemic Possibility," *Philosophical Review*, C, (1991): 581–605; Patrick Rysiew, "The Context-Sensitivity of Knowledge Attributions," *Nous*, XXXV, 4 (2001): 477–514; Jason Stanley, "Fallibilism and Concessive Knowledge Attributions," *Analysis*, LXV, 2 (2005): 126–31; Trent Dougherty and Patrick Rysiew, "Fallibilism, Epistemic Possibility, and Concessive Knowledge Attributions," *Philosophy and Phenomenological Research*, LXXVIII, 1 (2009): 123–132; Littlejohn, "Concessive Knowledge Attributions and Fallibilism," *op. cit.*; Charity Anderson, "Fallibilism and the Flexibility of Epistemic Modals," *Philosophical Studies*, CLXVII, 3 (2014): 597–606.

[7]The modal base is usually defined instead as a function from a world to a set of propositions (see Angelika Kratzer, "The Notional Category of Modality," in Eikmeyer and Rieser, eds., *Words, Worlds, and Contexts: New Approaches to Word Semantics* (Berlin: Walter de Gruyter, 1981), pp. 38–74). But this difference does not affect the arguments at issue here.

3 What is Possibly False Knowledge?

Worsnip's idea of QDR seems exactly right, and his objection to KPL contains important insights. What I wish to dispute is how all of this bears on whether there is possibly false knowledge.

In the previous section, we defined both KPL and PFK in terms of truth at a context:

> KPL (repeated here): For all contexts c, ⌜it is possible that not-ϕ⌝ is true at c only if ⌜I do not know that ϕ⌝ is true at c.
>
> Definition 1 (repeated here):
> PFK: there exists some context c such that ⌜I know that ϕ⌝ is true at c, and ⌜it is possible that not-ϕ⌝ is true at c.

These definitions come closest to matching Worsnip's text. But the definitions do not, I think, capture the real force of his arguments or bring out what is most interesting about (*)-sentences. Let me explain.

If we follow the above definitions, then Worsnip's argument for PFK is inconclusive for a very uninteresting reason. The reason is simply that there are now numerous accounts of epistemic modal discourse on which the notion of truth at a context is undefined for epistemically modalized sentences. The views I have in mind are so-called nonfactualist accounts of epistemic modality. Nonfactualist accounts deny that asserting a bare, epistemically modalized sentence functions to communicate a possible-worlds proposition, or even a centered-worlds proposition.[8] Instead, the function of asserting ⌜it is possible that ϕ⌝ is to call attention to ϕ's compatibility with the conversational common ground, to propose to make the common ground so compatible, or to propose a constraint on the credences of conversational participants.[9] On all of these theories, the function of epistemic modal discourse is not to describe what the world is like or to locate oneself in a world

[8]Bare epistemic modals are those used without a restrictor phrase: e.g. ⌜it is possible that ϕ⌝, as compared with ⌜for all I know, it is possible that ϕ⌝.

[9]See Eric Swanson, *Interactions with Context*, PhD thesis, Massachusetts Institute of Technology (2006); Eric Swanson, "How Not to Theorize about the Language of Subjective Uncertainty," in Andy Egan and Brian Weatherson, eds., *Epistemic Modality*, (Oxford: Oxford University Press, 2011), pp. 249–269; Eric Swanson, "The Application of Constraint Semantics to the Language of Subjective Uncertainty," *Journal of Philosophical Logic*, XLV, 2 (2016): 121–146; Yalcin, "Epistemic Modals," *op. cit.*; Seth Yalcin, "Nonfactualism about Epistemic Modality," in Andy Egan and Brian Weatherson, eds., *Epistemic*

or a context. Consequently, all of these theories will deny that the notion of truth at a context is well defined for epistemically modalized sentences.[10]

However, if truth at a context is not defined for epistemically modalized sentences, then there cannot be a context at which ⌜it is possible that not-ϕ⌝ is true, and thus PFK is false, according to Definition 1. Nonfactualism also raises a problem for Worsnip's argument against KPL. If truth at a context is undefined for epistemic modals, then KPL is trivially true.

I thus proceed on the assumption that Definition 1 is not the right way of capturing the relevant notion of possibly false knowledge and that KPL must also be reformulated.

Before setting out what I think is the right way to understand Worsnip's arguments, let me quickly review two reformulations that I do not find promising. On the first, we define PFK (and KPL) not in terms of truth at a context but in terms of truth at a point of evaluation. Nonfactualists often allow that the latter is defined for epistemic modals even if the former is not. For example, consider Seth Yalcin's definition of truth at a point of evaluation for epistemic "might"/"possible":[11]

$[\![\Diamond\phi]\!]^{c,s,w}$ is true iff $\exists w' \in s : [\![\phi]\!]^{c,s,w'}$ is true.

s is the so-called information state parameter, which ranges over *sets* of worlds—that is, information states. On this semantics, ⌜$\Diamond\phi$⌝ essentially determines a property of information states—being compatible with ϕ—as opposed to a property of a world, as on traditional, contextualist semantics.

We can then define PFK as follows:

> Definition 2:
> PFK: there exists some point of evaluation <c,s,w> such that $[\![K\phi]\!]^{c,s,w}$ is true and $[\![\Diamond\neg\phi]\!]^{c,s,w}$ is true, where ⌜$K\phi$⌝ abbreviates ⌜I know that ϕ⌝.

Modality, (Oxford: Oxford University Press, 2011), pp. 295–332; and Sarah Moss, *Probabilistic Knowledge* (Oxford: Oxford University Press, 2018). Other defenses of nonfactualism include Frank Veltman, "Defaults in Update Semantics," *Journal of Philosophical Logic*, xxv, 3 (1996): 221–261; Daniel Rothschild, "Expressing Credences," *Proceedings of the Aristotelian Society*, cxii, (1pt.1) (2012): 99–114; and Malte Willer, "Dynamics of Epistemic Modality," *Philosophical Review*, cxxii, 1 (2013): 45–92.

[10] For further discussion see Yalcin, "Epistemic Modals," *op. cit.* and Yalcin, "Nonfactualism about Epistemic Modality," *op. cit.*

[11] Yalcin, "Epistemic Modals," *op. cit.*

Given this definition, it is easy to show that PFK is true. However, many of the witnessing points of evaluation are clearly irrelevant to the existence of an epistemologically interesting notion of possibly false knowledge. For example, consider a point of evaluation where $s =$ the set of worlds compatible with an arbitrary smallest not-ϕ-compatible subset of the propositions in the speaker of c's total evidence in c at w. Here it follows that PFK is true as long as one can know that ϕ even though some tiny subset of one's total evidence happens to be compatible with not-ϕ.[12]

It is also implausible to define PFK by simply using epistemic modals:

> Definition 3:
> PFK: $\Sigma\phi$ $\exists S$ such that S knows that ϕ and it is possible that not-ϕ.[13]

Definition 3 employs the substitutional quantifier Σ: PFK is true iff there is some true substitution-instance of the substitutional quantifier scope. However, there is a problem here even if one employs a notion of "true" that is not equivalent to truth at a context or truth at a point of evaluation. The problem is that Definition 3 makes the truth of PFK hostage to epistemologically irrelevant facts about the interaction between epistemic modals and conjunction. For example, Mandelkern argues that something like QDR operates also at the level of semantic composition.[14] He proposes that the semantic value of an epistemic modal is defined only if the modal is interpreted relative to the information contained in the modal's local context. Since the local context of the right conjunct of a conjunction plausibly includes the information contained in the left conjunct, $\ulcorner\phi$ and it is possible that not-$\phi\urcorner$ must either be false or undefined. Similarly, if Mandelkern's theory is correct, then all substitution-instances of the substitutional quantifier scope in Definition 3 will also be either false or undefined, since knowledge is factive.

[12]Cf. Jessica Brown, *Fallibilism: Evidence and Knowledge* (Oxford: Oxford University Press, 2018) at pp. 174–175. One might try to solve this problem by taking the relevant points of evaluation for verifying PFK to be those consisting of a context and the index determined by that context. But the idea behind Yalcin's nonfactualism is precisely that there is no such thing as s_c. See Yalcin, "Epistemic Modals," *op. cit.*, Yalcin, "Nonfactualism about Epistemic Modality," *op. cit.*, and our objection to Definition 1 above.

[13]Some theorists might wish to modify the second conjunct so that it reads: $\ulcorner$it is possible for S that not-$\phi\urcorner$. But I concur with Worsnip, "Possibly False Knowledge," *op. cit.*, p. 226n. 3 that $\ulcorner$possible for S$\urcorner$, in the relevant sense, is not part of natural language epistemic modal discourse.

[14]Mandelkern, "Bounded Modality," *op. cit.*

The moral here is that PFK must be formulated in the meta-language, but not in terms of truth at a context or truth at a point of evaluation. I suggest, then, that we think of PFK and KPL as views about consequence or entailment relations among sentences in natural language, where such entailment relations are represented by $\models$. Here is my proposed reformulation of KPL and three possible definitions of PFK:

KPL2: $\Diamond\neg\phi \models \neg K\phi$

Definition 4:
PFK: $(K\phi \wedge \Diamond\neg\phi) \not\models \bot$

Definition 5:
PFK: $K\phi \not\models \neg\Diamond\neg\phi$

Definition 6:
PFK: $K\phi \not\models \neg\Diamond\neg\phi$ and $\Diamond\neg\phi \not\models \neg K\phi$

Definitions 4-6 avoid the problems we raised above for definitions 1 and 2. That is, definitions 4-6 neither presuppose that nonfactualism is false nor entail that the mere existence of some point of evaluation at which $\ulcorner K\phi \urcorner$ and $\ulcorner \Diamond\neg\phi \urcorner$ are true verifies PFK. Definition 4 employs a CKA and faces the same problem as Definition 3 regarding conjunction. So I propose that either Definition 5 or Definition 6 is the best formulation of PFK. I will employ Definition 5 for ease of exposition, but the choice between definitions 4-6 will not affect my arguments below.

Worsnip's argument against KPL2 and for PFK should then be understood as follows:

(*) (repeated here) I think I know that ϕ, but it is possible that not-ϕ.

P1. (*)-sentences are felicitous.

P2. $(\Diamond\phi \wedge \neg\phi)$ is infelicitous.

P3. Asserting a (*)-sentence involves commitment to $\ulcorner(\Diamond K\phi \wedge \Diamond\neg\phi)\urcorner$.

P4. $(\Diamond K\phi \wedge \Diamond\neg\phi) \not\models (\Diamond K\phi \wedge \neg K\phi)$ (by P1, P2, and P3)

P5. $\Diamond\neg\phi \not\models \neg K\phi$ [KPL2 is false] (by P4)

C. $K\phi \not\models \neg\Diamond\neg\phi$ [PFK is true] (by P5 and Definition 5)

Let us reflect on Worsnip's argument. To begin, his case against KPL2 is plausible. In fact, we can strengthen his argument by avoiding reliance on P2.[15] Notice that if KPL2 is correct, then if we assume $\models$ is monotonic and validates conjunction introduction, we have:

$$(\Diamond K\phi \wedge \Diamond\neg\phi) \models (\neg K\neg K\phi \wedge \neg K\phi)$$

Hence, by P3, asserting a (*)-sentence would involve commitment to a Moore-paradoxical sentence of the form $\ulcorner(\neg K\phi \wedge \phi)\urcorner$. But even if one finds (*)-sentences slightly degraded, as Keith Derose reports,[16] (*)-sentences clearly sound far better than Moore-paradoxical sentences. So KPL2 must be false.

It would then seem to be a short step from the failure of KPL2 to the truth of PFK:

P5. (repeated here) $\Diamond\neg\phi \not\models \neg K\phi$ [KPL2 is false]

C. (repeated here) $K\phi \not\models \neg\Diamond\neg\phi$ [PFK is true]

Surprisingly, however, the inference from P5 to C is not at all straightforward. This inference is valid if the consequence relation for natural language sentences—$\models$—respects the entailment patterns of classical logic. But many authors have advanced alternative, non-classical consequence relations for natural language sentences. As we will see below, these non-classical consequence relations allow one to reject KPL2, embrace QDR, and predict the felicity of (*)-sentences all while denying PFK, as defined by definitions 4-6. Thus, it turns out that Worsnip's argument does not succeed in establishing the existence of possibly false knowledge, since one can accept all of his premises while denying his conclusion. Or: Worsnip's argument is valid but not compelling since it relies on the unargued and now-controversial assumption that the consequence relation appropriate to natural language is classical.

[15]There is some controversy in the literature over P2, as compared with the claim that $\ulcorner(\neg\phi\wedge\Diamond\phi)\urcorner$ is infelicitous. For discussion see Bob Beddor and Simon Goldstein, "Believing Epistemic Contradictions," *Review of Symbolic Logic*, 1, (2018): 87–114, at §8.1.

[16]Keith DeRose, "Contextualism and Fallibilism," in J. Ichikawa, ed., *The Routledge Handbook of Epistemic Contextualism* (London: Routledge, 2017) at n. 19.

One example of a non-classical consequence relation with the features promised above comes from Yalcin's work. He first defines the notion of an information state's *incorporating* a sentence ϕ in a context c, ϕ_c:[17]

> s incorporates ϕ_c iff $\forall w \in s : [\![\phi]\!]^{c,s,w}$ is true.

Intuitively, an information state incorporates an ordinary, non-modalized sentence just in case the proposition expressed by that sentence in c is true at all of the worlds in the information state. An information state incorporates a modalized sentence just in case the property of information states expressed by the modal is true of the information state. For example, an information state s incorporates $\ulcorner \Diamond \phi \urcorner$ in c just in case s has the property of being compatible with ϕ_c.

Yalcin uses this notion of incorporation to define a consequence relation—$\models_I$—he calls *informational consequence*:

> A set of sentences Γ is such that $\Gamma \models_I \phi$ iff for every context c and information state s, if s incorporates ψ_c for every $\psi \in \Gamma$, then s incorporates ϕ_c.

On informational consequence, valid arguments preserve incorporation in every context: in every context, every information state that incorporates the premises in that context incorporates the conclusion in that context.[18]

Informational consequence rejects PFK:

[17] I follow Justin Bledin, "Logic Informed," *Mind*, CXXIII, 490 (2014): 277–316 in replacing Yalcin's talk of "acceptance" with "incorporation".

[18] For defense of related notions of consequence in dynamic semantics see Veltman, "Defaults in Update Semantics," *op. cit.*; Anthony S. Gillies, "On Truth-Conditions for If (but Not Quite Only If)," *Philosophical Review*, CXVIII, 3 (2009): 325–349.; Willer, "Dynamics of Epistemic Modality," *op. cit.*; and Malte Willer, "A Remark on Iffy Oughts," *Journal of Philosophy*, CIX, 7 (2012): 449–461. See also Kolodny and MacFarlane's notion of *quasi-validity* (Niko Kolodny and John MacFarlane, "Ifs and Oughts," *Journal of Philosophy*, CVII, 3 (2010): 115–143.). Bledin, "Logic Informed," *op. cit.* defends informational consequence as a general account of logical consequence. See also Moritz Schulz, "Epistemic Modals and Informational Consequence," *Synthese*, CLXXIV, 3 (2010): 385–395; Wesley H. Holliday and Thomas Icard, "Indicative Conditionals and Dynamic Epistemic Logic," in Jerome Lang, ed., *Proceedings of the Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge* (Liverpool: TARK, 24-26 July 2017), pp. 337–351; Justin Bledin and Tamar Lando, "Closure and Epistemic Modals," *Philosophy and Phenomenological Research*, XCVII, 1 (2018): 3–22; Simon Goldstein,"Epistemic Modal Credence," *Philosophers' Imprint*, (forthcoming); and Paolo Santorio, "Trivializing Informational Consequence," *Philosophy and Phenomenological Research*, (forthcoming).

$$K\phi \models_I \neg\Diamond\neg\phi$$

This result follows from the factivity of knowledge: if an information state incorporates $\ulcorner K\phi \urcorner$, the information state must be incompatible with $\ulcorner \neg\phi \urcorner$ and must therefore incorporate $\ulcorner \neg\Diamond\neg\phi \urcorner$. We also have $(K\phi \wedge \Diamond\neg\phi) \models_I \bot$ for essentially the same reason. In other words, if informational consequence is correct, then CKAs are contradictory.[19]

What is more surprising, however, is that informational consequence allows us to reject PFK while joining Worsnip in denying KPL2. Notice that any plausible semantics for K yields the following:

$$\Diamond\neg\phi \not\models_I \neg K\phi$$

KPL2 fails as long as it is possible for an information state to be undecided about both whether ϕ and whether $K\phi$.[20] That is, an information state might be compatible with both ϕ and $\neg\phi$ and hence incorporate $\ulcorner \Diamond\neg\phi \urcorner$ while being compatible with both $K\phi$ and $\neg K\phi$ and hence failing to incorporate $\ulcorner \neg K\phi \urcorner$. More generally, if one were to propose a semantics for K on which $\Diamond\neg\phi \models_I \neg K\phi$, then we would have the absurd result that $\Diamond K\phi \models_I \phi$.[21] Note also that informational consequence is consistent with QDR and predicts that (*)-sentences are felicitous.[22] The same goes for related non-classical consequence relations in dynamic semantics that also allow one to reject both KPL2 and PFK.[23]

It will be helpful to see where exactly Worsnip's reasoning goes wrong according to informational consequence. Worsnip writes the following about a speaker's utterance of a (*)-sentence:

---

[19] Bob Beddor, "New Work for Certainty," *Philosophers' Imprint*, xx, 8 (2020): 1–25, at p. 20 makes a similar observation in the context of dynamic semantics. Note that informational consequence also yields $(\phi \wedge \Diamond\neg\phi) \models_I \bot$ and $\phi \models_I \neg\Diamond\neg\phi$.

[20] Formally, a countermodel requires only that there exist some context c and information state s such that $\exists w \in s : [\![\phi]\!]^{c,s,w}$ is false and $\exists w \in s : [\![K\phi]\!]^{c,s,w}$ is true.

[21] *Proof.* $\Diamond\neg\phi \models_I \neg K\phi$ iff $\forall c \forall s(\exists w \in s : [\![\phi]\!]^{c,s,w}$ is false only if $\forall w \in s : [\![\neg K\phi]\!]^{c,s,w}$ is true) iff $\forall c \forall s(\exists w \in s : [\![K\phi]\!]^{c,s,w}$ is true only if $\forall w \in s : [\![\phi]\!]^{c,s,w}$ is true) iff $\Diamond K\phi \models_I \phi$ $\square$.

[22] Informational consequence predicts that (*)-sentences are felicitous since $(\Diamond K\phi \wedge \Diamond\neg\phi) \not\models_I \bot$ given the assumptions about K made in n. 20. Informational consequence entails QDR on the assumption that assertions are proposals to make the context set of the common ground incorporate the asserted sentence in the context in which it is asserted.

[23] See Veltman, "Defaults in Update Semantics," *op. cit.*, Willer, "Dynamics of Epistemic Modality," *op. cit.*, and Willer, "A Remark on Iffy Oughts," *op. cit.*

> [I]f her utterance is felicitous, then it is also still a possibility . . . that she does know that P (again, even at standards consistent with a context of utterance where she can truly say "it is possible that not-P"). And if that possibility does obtain—in this case, if she does know [that P]—then she knows this compatibly with the epistemic possibility that [not-P]. She has possibly false knowledge.[24]

According to informational consequence, the problem lies in Worsnip's conditional: "if she does know [that P,] then she knows this compatibly with the epistemic possibility that [not-P]." On informational consequence, if one supposes that $K\phi$, then one cannot at the same time coherently suppose that $\Diamond\neg\phi$, since no coherent information state incorporates both sentences. And this is so even though one can initially suppose that $\Diamond\neg\phi$ or $(\Diamond K\phi \wedge \Diamond\neg\phi)$ without thereby supposing that $\neg K\phi$.[25] Essentially, the idea here is that an information state can be undecided as to whether $K\phi$ and whether ϕ, but once the information state becomes decided as to $K\phi$, it cannot remain undecided as to whether ϕ, since knowledge is factive. Hence, on informational consequence, the idea of possibly false knowledge is actually incoherent.[26]

Now, we've seen that Worsnip's argument for PFK is inconclusive. But I wish to emphasize that his argument contains an important lesson that should not be overlooked. Worsnip's argument shows that the felicity of (*)-sentences forces those who accept a classical consequence relation to admit the existence of possibly false knowledge. We can express this insight in terms of an inconsistent triad:

(1) $(\Diamond K\phi \wedge \Diamond\neg\phi) \not\models (\neg K\neg K\phi \wedge \neg K\phi)$

[24]Worsnip, "Possibly False Knowledge," *op. cit.*, at pp. 238–239.

[25]Given the assumptions about K made in n. 20, one can even coherently suppose $\Diamond(K\phi \wedge \Diamond\neg\phi)$ even though one cannot coherently suppose $(K\phi \wedge \Diamond\neg\phi)$. We also have: $\not\models_I \neg(K\phi \wedge \Diamond\neg\phi)$. But the latter holds iff $\Diamond(K\phi \wedge \Diamond\neg\phi) \not\models_I \bot$, so I take it that $\not\models_I \neg(K\phi \wedge \Diamond\neg\phi)$ does not establish the existence of possibly false knowledge.

[26]Worsnip comes closest to addressing informational consequence when he observes that $K\phi$ and $\Diamond\neg\phi$ are what he calls *dynamically contradictory*: one cannot assert them together without contradicting oneself (Worsnip, "Possibly False Knowledge," *op. cit.*, at pp. 239–240). He seems to argue that this notion of contradiction/consequence cannot be the correct one for settling the existence of possibly false knowledge, since if it were, then $\ulcorner\Diamond\neg\phi\urcorner$ would entail $\ulcorner\neg\phi\urcorner$ since $\ulcorner(\phi \wedge \Diamond\neg\phi)\urcorner$ is also dynamically contradictory. But note that informational consequence yields $(\phi \wedge \Diamond\neg\phi) \models \bot$ but not $\Diamond\neg\phi \models \neg\phi$ (see Yalcin, "Epistemic Modals," *op. cit.*).

(2) $K\phi \models \neg\Diamond\neg\phi$

(3) $\models$ is classical.

This inconsistent triad places great pressure on theorists who endorse (2) and (3). Such theorists must deny (1) and thus face the difficult task of explaining why (*)-sentences do not sound Moore-paradoxical. Alternatively, if one grants that asserting a (*)-sentence does not involve Moore-paradoxical commitment, then if one wishes to deny possibly false knowledge, one must embrace nonclassicality.

4 Possibly False Knowledge and Fallibilism

Worsnip describes PFK as a robust version of fallibilism, as compared with the "less ambitious" fallibilism he finds in the work of Jason Stanley, who denies PFK.[27] It is worth reflecting, then, on whether and to what extent fallibilists are indeed committed to PFK.

The discussion of the previous section reveals that commitment to PFK is more weighty than I think most epistemologists have realized. Maintaining PFK requires not just denying the orthodox semantics on which epistemic modals describe knowledge. One must also reject several non-classical accounts of consequence that have risen to prominence in recent years. The previous section also shows that if informational consequence is correct, then CKAs are contradictory. Thus, one cannot establish the consistency of CKAs simply by giving a pragmatic explanation of their infelicity and rejecting the knowledge-describing semantics for epistemic modals. Maintaining the consistency of CKAs requires denying informational consequence and related notions of consequence in dynamic semantics.

Some fallibilists may be willing to pay this price: the epistemological merits of fallibilism are so great that if some particular semantics gets in the way, then so much the worse for the semantics.[28] But I find another

[27]See Worsnip, "Possibly False Knowledge," *op. cit.*, p. 225; *ibid.*, p. 225n. 1; and Stanley, "Fallibilism and Concessive Knowledge Attributions," *op. cit.* Matthew Benton says much the same about a view like Stanley's (see Matthew A. Benton, "Knowledge, Hope, and Fallibilism," *Synthese*, online first (2018): 1–17). Baron Reed argues that Stanley's view is inconsistent with fallibilism (see Baron Reed, "Fallibilism, Epistemic Possibility, and Epistemic Agency," *Philosophical Issues*, XXIII, 1 (2013): 40–69, at pp. 52–53).

[28]Cf. Reed, "Fallibilism, Epistemic Possibility, and Epistemic Agency," *op. cit.* on Stanley's semantics.

inference more attractive: once one sees what is involved in maintaining PFK or the consistency of CKAs, one ought to reassess why exactly one thought fallibilists were committed to these theses in the first place.

Fallibilism is often roughly characterized as the view that knowledge is compatible with the possibility of error. But why think that "possibility" in the relevant sense is the same sense of "possibility" expressed by natural language epistemic modals?[29] The discussion of the previous section should make it clear why it is at best a risky proposition to simply *define* fallibilism as a view about the relation between "knows" and natural language "might"/"possible". If one gives such a definition in terms of truth at a context, then fallibilism is refutable by nonfactualism about epistemic modals. If one gives such a definition in terms of truth at a point of evaluation, then fallibilism is too cheap to be of epistemological interest. If one gives such a definition by using epistemic modals, then fallibilism is refutable by Mandelkern's semantics. If one gives such a definition in terms of consequence, then fallibilism is refutable by informational consequence.

I suggest, then, that we abandon the idea that natural language talk of epistemic possibility is somehow built into the very concept of fallible knowledge. Fallibilism is the view that knowledge is compatible with the possibility of error, but the relevant sense of "possibility" can only be settled by substantive epistemological theorizing, not the semantics of epistemic modals. It may turn out that "possibility" in the relevant sense coincides with the meaning of natural language talk of what is epistemically possible—for example, if the fallibilism-relevant sense of "possibility" is compatibility with one's evidence and natural language epistemic modals express this same sense of compatibility.[30] But it equally may not turn out this way—for example, if the fallibilism-relevant sense of "possibility" is compatibility with one's evidence but some version of nonfactualism is true for natural language talk of "possibility".

[29]Stanley and Wesley Holliday press this same question; see Stanley, "Fallibilism and Concessive Knowledge Attributions," *op. cit.* and Wesley H. Holliday, "Fallibilism and Multiple Paths to Knowledge," *Oxford Studies in Epistemology*, v, (2015): 97–144 at n. 15.

[30]For a view like this see Dougherty and Rysiew, "Fallibilism, Epistemic Possibility, and Concessive Knowledge Attributions," *op. cit.* and Trent Dougherty,"Fallibilism," in Duncan Pritchard and Sven Bernecker, eds., *The Routledge Companion to Epistemology* (London: Routledge, 2011). As they recognize, such a view can vindicate fallibilism only if E = K is false.

The upshot is that there is no straightforward way to use linguistic data about natural language epistemic modals to either verify or refute the fallibilist thesis, as Worsnip and several others have tried to do.[31] Any successful argument would have to proceed in two stages. One must first identify the sense of "possibility" relevant to the truth of fallibilism. One must then show that the natural language epistemic modal sentences constituting one's linguistic data express this exact same sense of "possibility". Stage one is complicated by the fact that one can neither use nor mention epistemic modals to simply define the fallibilism-relevant sense of "possibility". Stage two is complicated by the fact that the meaning of epistemic modals is hotly disputed. It is no surprise, then, that fallibilism remains controversial.

[31]See Dylan Dodd, "Against Fallibilism," *Australasian Journal of Philosophy*, LXXXIX, 4 (2011): 665–685; Benton, "Knowledge, Hope, and Fallibilism," *op. cit.*; and perhaps David Lewis, "Elusive Knowledge," *Australasian Journal of Philosophy*, LXXIV, 4 (1996): 549–567.