Synthese (2021) 198:3855–3871
https://doi.org/10.1007/s11229-019-02314-6

Knowledge-first believing the unknowable

Simon Wimmer[1]

Received: 20 June 2018 / Accepted: 2 July 2019 / Published online: 10 July 2019

Abstract
I develop a challenge for a widely suggested knowledge-first account of belief that turns, primarily, on unknowable propositions. I consider and reject several responses to my challenge and sketch a new knowledge-first account of belief that avoids it.

Keywords Knowledge-first · Belief · Philosophy of mind · Epistemology

1 Introduction

Sometimes we believe p without knowing p. For example, Jess believes that Aristotle is Plato—something she does not and cannot know, since it is metaphysically impossible that Aristotle is Plato.

A *belief-first* view of knowledge predicts this datum. On such a view, knowing p is something like the conjunction of believing p with further factors, e.g. truth. As one of these factors, namely truth, does not obtain in Jess' case, she believes without knowing. However, specifying the further factors posited by the belief-first view has proven difficult.[1] Given this, we might adopt a *knowledge-first* account instead and use knowing p or one of its distinctive properties as a primitive in theorizing about belief.[2]

What might a knowledge-first view of belief look like? Because we sometimes believe p without knowing p, believing p cannot be knowing p or something like the

[1] Williamson (2000) makes much of this difficulty. Several chapters in Greenough and Pritchard (2009) discuss Williamson's argumentative strategy. For further considerations against belief-first views see, e.g., Holton (2017), Hyman (2015, 2017), Nagel (2013), and Vendler (1972). See Butterfill (2013), Roessler (2013), Rose (2015), Rysiew (2013), McGlynn (2017) for discussion of Nagel's, and Dunn and Suter (1977), Jones (1975), Rosenthal (1976) for discussion of Vendler's argument. I here remain neutral on whether the considerations just cited should move us to abandon belief-first views.

[2] This does not require that we treat knowing p or one of its distinctive properties as primitive *simpliciter*. Going knowledge-first, in the intended sense, is consistent with understanding knowledge in terms of an ability, as Hyman (2015) does, for instance.

✉ Simon Wimmer
simon.wimmer@gotonet.at

[1] Department of Philosophy, Social Sciences Building, University of Warwick, Coventry CV4 7AL, UK

conjunction of knowing p with further factors. A disjunctive view, on which believing p is something like the disjunction of knowing p and opining p (or some suitable alternatives) seems problematic too. For how are the disjuncts other than knowing p to be understood? If they are understood as merely believing p, for instance, the account is circular.[3]

A more promising option models a knowledge-first view on *truth-first* views of belief familiar from, e.g., Braithwaite (1932) and Marcus (1990).[4] These views say, roughly, that to believe p is to Φ, for instance, be disposed to act, as if p were true.[5] On the corresponding knowledge-first view (henceforth 'OLD'), to believe p is to Φ as if p were known, i.e. as if one knew p. Because we sometimes Φ, e.g., are disposed to act, as if we knew p without knowing it, OLD predicts that we sometimes believe p without knowing it.

A view like OLD is suggested in various places. For instance, Williamson (2000) says that "believing p is, roughly, treating p as if one knew p" (p. 47).[6] Similarly, Jennifer Nagel (2017) writes that "Believing is [...] a shadow or after-effect of knowing: the deceived agent who reaches for the basket behaves as if she knew that the ball is in the basket" (p. 537).[7] Finally, John Hyman (2017) proposes that "[...] we can define the belief that p [...] as the disposition to act (think, feel) as one would if one knew that p." (p. 284)[8,9]

I take on two tasks here. First, I develop a challenge for OLD that turns, primarily, on unknowable propositions.[10] Roughly, I argue that, depending on what further commitments we take on, the view either under- or overgenerates, predicting that one believes either too few or too many unknowable propositions. Second, I sketch a sharpened knowledge-first view (henceforth 'NEW') for which unknowable propositions raise no special difficulties. On this view, to believe p is to Φ with respect to p as one would with respect to some knowable proposition q if one knew q.

Since NEW avoids under- and overgenerating in the way OLD does, it is a better starting point for assessing the prospects of going knowledge-first about belief. However, whether it also yields correct predictions in all cases, and so is ultimately plausible, is not my concern here. This is because for NEW to make predictions in all cases, we

[3] See Williamson (2000, pp. 42–6) for further discussion.

[4] Hyman (2017) compares truth- and knowledge-first views.

[5] 'Φ' will be used as a placeholder, rather than variable, throughout.

[6] Holton (2017) and Nagel (2013) seem to endorse this suggestion. Williamson, however, only endorses a refinement of this suggestion. See Sect. 2 for brief discussion of that refinement. See Roessler (2013, pp. 323–4) for brief critical discussion of both the suggestion in the text and Williamson's refinement.

[7] Note, however, that Nagel does not intend her suggestive remark to put forward an account of belief.

[8] Hyman (2017, p. 285) suggests that, if we take knowledge not to entail belief, but only to be normally accompanied by it, "we can [instead] define the belief that p as the disposition to act (think,feel) as one *normally* or *generally* would if one knew that p." My challenge below arises *mutatis mutandis* also for this view.

[9] If 'think,' on one reading, is synonymous with 'believe' (e.g. Hawthorne et al. 2016), we may worry that Hyman's appeal to thinking renders his definition circular. However, there is a reading of 'think' on which it denotes occurrences of thinking, conscious processes, rather than states of thinking. On this reading, Hyman's definition is not circular.

[10] Hyman (2017, p. 286) suggests that unknowable propositions may be problematic for his view.

must substitute for 'Φ', interpret 'Φing *with respect to* p', and do much else. But these are tasks I leave for another occasion.

Section 2 makes OLD and NEW more precise. Section 3 shows that, given two orthodox assumptions, OLD overgenerates. Section 4 argues that rejecting the first assumption lacks independent motivation; Sect. 5 that even if we reject the second assumption, OLD overgenerates. Finally, Sect. 6 shows that NEW avoids both under- and overgenerating in the way OLD does.

2 Regimentations

I now regiment NEW and OLD. This allows me to state my challenge for OLD, and explain why NEW does not face it, particularly clearly.

Start with OLD. The clause to the right of 'as', 'if one knew it,' is in the subjunctive mood. We may thus interpret it as the antecedent of a would-counterfactual whose consequent is elided.[11] Recovering the elided material, OLD reads: to believe p is to Φ as one would Φ if one knew it.

I adopt a semantics of the comparative particle 'as' inspired by Bcking (2017)'s treatment of German 'wie'.[12] On this semantics, 'as' denotes an equivalence relation between individuals or eventualities. This relation holds of two individuals or eventualities just in case they are the same in a certain contextually salient way.[13] One Φs as one would Φ if one knew it, then, just in case one Φs in some contextually salient way F and if one knew it, one would Φ in F.

I now regiment old as per:

OLD For all x and p, x believes p iff: for some contextually salient way F, (i) x Φs in F and (ii) if x knew p, x would Φ in F.

Regimenting new in parallel fashion, we get:

NEW For all x and p, x believes p iff: for some knowable q and contextually salient way F, (i) x Φs with respect to p in F and (ii) if x knew q, x would Φ with respect to q in F.[14]

My regimentations highlight three important differences between OLD and NEW. First, NEW, but not OLD, is strict. Only NEW restricts the domain of one of its quantifiers,

[11] I assume that OLD's counterfactual is a would-counterfactual. This is no trivial assumption. But since Hyman, in the quotation above, explicitly appeals to a would-counterfactual, it seems reasonable for present purposes. Moreover, a challenge analogous to mine arises even if the would-counterfactual is replaced with a might-counterfactual.

[12] Bücking in turn draws on Umbach and Gust (2014).

[13] What determines which ways are salient in a context? For present purposes, I remain neutral on this question, but some (non-exclusive) options are: ostension, the question(s) under discussion, pragmatic presuppositions made by discourse participants, and charity constraints. My challenge arises on any of these options.

[14] My regimentations of OLD and NEW diverge from the unregimented views in one respect: their modal force. The unregimented views define what it is to believe a proposition. Such definitions are not equivalent to a biconditional. However, they entail corresponding biconditionals. See Correia (2017), Dorr (2016), Rosen (2015). Consequently, the unregimented views entail the regimented ones. That suffices for present purposes.

viz. that of the existential quantifier binding q, to all and only knowable propositions. Second, NEW, but not OLD, is flexible. OLD appeals to one proposition variable bound by a universal quantifier only. By contrast, NEW appeals to two proposition variables, one bound by a universal, the other by an existential quantifier. Finally, NEW, but not OLD, relativizes Φing to a proposition. As we will see in Sect. 6, all three differences play a role in making NEW preferable to OLD.

OLD and NEW may also be regimented in other ways. For instance, we might not require sameness but only similarity between ways of Φing, might appeal to the fully specific way in which x Φs (with respect to p) instead of a contextually salient one, or might replace the would-counterfactual with a normality or strict conditional. However, although I do not show this here for reasons of space, no plausible regimentation of OLD avoids my challenge.

One comment before we turn to my challenge for OLD. NEW resembles a proposal by Williamson. In the passage I quoted in the introduction, Williamson uses the adverbial modifier 'roughly.' The suggestion contained in that passage is an approximation only. Williamson refines it thus:

> [...] to believe p is to treat p as if one knew p—that is, to treat p in ways similar to the ways in which subjects treat propositions which they know. (2000, pp. 46–47)

Much like NEW, this view appeals to two proposition variables, p and that introduced by 'propositions'. Williamson does not say why he offers his refinement. My challenge to OLD provides a clue. OLD is not flexible enough. Williamson's refinement is. How exactly it avoids my challenge, however, is an issue for another occasion.

3 The initial challenge

I now argue that, given two orthodox assumptions, OLD overgenerates. It predicts that, in suitable contexts, everyone believes all unknowable propositions. For concreteness, I focus on the version of OLD we get by substituting 'treats p' for 'Φ.' As is easily verified, though, versions of OLD inspired by the quotations from Nagel and Hyman face the same challenge.

3.1 Two assumptions

My first assumption is that at least one proposition is metaphysically impossible to know (henceforth 'cannot be known' or 'is unknowable'). This assumption is entailed by plausible claims about knowledge. Knowledge (strictly) entails truth.[15] More precisely, for every metaphysically possible world w, subject x and proposition p, if x knows p in w, p is true in w. By this claim, all metaphysically impossible propositions, i.e. all propositions that are true at no metaphysically possible world, are

[15] For brevity, I omit the modifier 'strictly' below.

unknowable.[16] For instance, assuming the necessity of distinctness, one cannot know the proposition that Aristotle is Plato.

Given other orthodox assumptions, the claim that knowledge entails truth also entails that many other metaphysically possible propositions are unknowable.[17] Take, e.g., the assumption that knowledge distributes over conjunction, according to which, for all metaphysically possible worlds w, subjects x, and propositions p and q, if x knows p and q in w, x knows p in w and x knows q in w. Given this assumption and the claim that knowledge entails truth, propositions like the metaphysically possible proposition that (Aristotle is the teacher of Alexander the Great and one does not know that Aristotle is the teacher of Alexander the Great) are unknowable for one.[18]

My second assumption is that would-counterfactuals with metaphysically impossible antecedents (henceforth 'counterpossibles') are vacuously true.[19] This assumption is entailed by the standard semantics of would-counterfactuals, inherited from Kratzer (1979), Lewis (1973), Stalnaker (1968).

On this semantics, a would-counterfactual 'If it were the case that ϕ, it would be the case that ψ' (henceforth also '$\phi \mathrel{\Box\!\!\rightarrow} \psi$') is true at a world w just in case all the metaphysically possible ϕ-worlds closest to w are ψ-worlds.[20] In effect, the operator '$\Box\!\!\rightarrow$' is a universal quantifier whose domain is restricted to the metaphysically possible ϕ-worlds closest to the world of evaluation. But if there is no metaphysically possible ϕ-world, the domain of '$\Box\!\!\rightarrow$' is empty. And if the domain of a universal quantifier heading a universally quantified sentence '$\forall x\, \phi(x)$' is empty, that sentence is vacuously true.[21] Thus, a would-counterfactual with a metaphysically impossible antecedent is vacuously true.[22]

[16] On a coarse-grained, possible worlds view of propositions, as in Stalnaker (1987), there is only one metaphysically impossible proposition. Whether we adopt such a coarse-grained or a more fine-grained view will not matter to our discussion.

[17] There may also be propositions that cannot be known for reasons not involving the entailment from knowledge to truth, though their status is more controversial. For instance, an epistemicist about vagueness like Williamson (1996) would hold that whether a borderline case of a vague concept F falls in the extension of that concept cannot be known since knowledge entails safety, that is, that x could not easily have falsely believed p, and one could easily have had false beliefs about whether a borderline case of a vague concept F falls in the extension of that concept. Since these considerations take us into controversial territory, I focus on less controversial examples.

[18] The argument for this is a variant of Fitch's paradox. Abbreviate 'Aristotle is the teacher of Alexander the Great' with 'α' and 'one does not know that α' with '$\neg K\alpha$.' Because knowledge distributes over conjunction, one's knowing that (α and $\neg K\alpha$) entails that one knows that α and that one knows that $\neg K\alpha$. But by the entailment from knowledge to truth, the second conjunct entails that $\neg K\alpha$. Yet, according to the first conjunct, one does know that α. Contradiction. To hold onto the assumptions that knowledge entails truth and distributes over conjunction, we must say that (α and $\neg K\alpha$) is unknowable for one. See Williamson (2000, ch.12) for discussion.

[19] For references see footnote 30.

[20] A ϕ-world is a world at which 'ϕ' is true.

[21] See, e.g., Barker-Plummer (2011).

[22] Note that the universal quantifier used in stating the truth-conditions of a would-counterfactual is not the English language determiner 'all.' Given this, one cannot avoid vacuous truth for counterpossibles by arguing, following Geurts (2008), that 'all' triggers an existential presupposition and so yields a truth-value gap if it is common ground that its domain is empty.

3.2 OLD overgenerates

Joan is a philosophy professor. She denies that Aristotle is Plato ('*a*') when asked whether *a*, refuses to rely on *a* in her reasoning, and so on. Joan's way of treating *a* strongly suggests that she does not believe it, even that she believes $\neg a$. However, given our two assumptions, we need only suppose that her way of treating *a* is contextually salient for OLD to predict that she believes *a*.

To see this, let Joan's actual way of treating *a* be *D*. Suppose also *D* is contextually salient, perhaps because we are pointing to Joan whilst she denies that *a* or ask whether Joan treats *a* in *D*. Joan satisfies OLD's first condition. But consider

1. If Joan knew *a*, Joan would treat *a* in *D*.

Because *a* is unknowable, there is no metaphysically possible world at which 1's antecedent is true. By our second assumption, 1 is thus true. Joan also satisfies OLD's second condition. OLD therefore predicts that Joan believes *a*.

Joan, *a* and *D* are an arbitrary subject, unknowable proposition and way of treating a proposition. So, we can generalize from Joan's case. For every subject x, unknowable proposition p_u and way of treating a proposition F, if x treats p_u in F and F is contextually salient, x believes p_u. Consequently, further counterexamples to OLD arise wherever a subject treats an unknowable proposition p_u in some way F, F is contextually salient, and treating p_u in F suggests that one suspends judgment about, does not believe, or disbelieves p_u.

The problem is compounded because ways of treating propositions are cheap. Plausibly, for every subject x and proposition p, x treats p in some way—even if for most p that way of treating it involves merely failing to consider p and similar negative occurrences. Since the unknowable propositions form a proper subset of the set of propositions, this means that for every subject x and unknowable proposition p_u, x treats p_u in some way. But then for every subject x, unknowable proposition p_u and some way of treating a proposition F, if F is contextually salient, x believes p_u.

One upshot of this is that, in contexts where the ways in which every subject treats each unknowable proposition are salient (perhaps because we ask how each subject treats each unknowable proposition), OLD predicts that every subject believes all unknowable propositions. In short, in suitable contexts, everyone believes all unknowable propositions.

3.3 Two responses

OLD's predictions may remind us of predictions made by possible-worlds models of propositions.[23] On such models, a proposition is the set of metaphysically possible worlds at which it is true. Thus, since metaphysically necessary propositions are true at all metaphysically possible worlds, all metaphysically necessary propositions are identified with the set of all metaphysically possible worlds. They are all the same proposition. Thus, believing some metaphysically necessary proposition entails believing all of them. If, as seems plausible, every subject believes at least one meta-

[23] See Stalnaker (1987) for a sympathetic discussion of such models.

physically necessary proposition, the possible worlds model of propositions predicts that every subject believes all of them.

Since both the possible-worlds model and OLD overgenerate, we may hope that responses given on behalf of the possible-worlds model generalize to OLD. But this is not so. OLD is even worse off than the possible-worlds model.

A common response on behalf of the possible-worlds model interprets it as an idealization.[24] To develop this response, we may say that the model idealizes by describing the 'equilibrium' state of rational believers. Perhaps, such a state is one ordinary believers are never in. But this is so only because of forces impeding their rationality. Absent these, ordinary believers would believe all metaphysically necessary propositions. A response of this kind is implausible if offered on behalf of OLD, however. That, in suitable contexts, everyone believes all unknowable propositions does not accurately describe the 'equilibrium' state of rational believers. In such a state they would generally avoid beliefs in unknowable propositions.

To develop the idealization response on behalf of the possible-worlds model, we may also say that the model describes the 'ideal' state of believers. Perhaps, ordinary believers are never in this state. But it would be ideal if they were. In some sense of 'ought', they ought to be in this state. A response of this kind too is implausible if offered on behalf of OLD. That, in suitable contexts, everyone believes all unknowable propositions does not accurately describe an 'ideal' state of believers. Other things being equal, one ought not to believe unknowable propositions.

Defenders of OLD cannot write off OLD's predictions as idealizations in the two ways sketched, even if parallel maneuvers help the possible-worlds model. Thus, defenders of OLD must block my argument.

Before considering whether rejecting one of my two assumptions is a promising way of doing so, consider a response that retains them. My argument assumed that unknowable propositions are in the domain of the universal quantifier binding OLD's proposition variable. So, to block my argument, defenders of OLD might restrict that domain to all and only knowable propositions.

However, doing so restricts OLD's purview. Given the restriction, OLD predicts that subjects cannot believe unknowable propositions. But, they can and do. Mathematicians and metaphysicians, for instance, often believe false propositions about their respective subject matter. Yet false propositions about mathematics and metaphysics alike are false in all metaphysically possible worlds, so metaphysically impossible, and so, by the entailment from knowledge to truth, unknowable.[25] Our account of belief, if it is to be fully general, should predict that we can and sometimes do believe (at least some) unknowable propositions. On the current response, OLD does not and so undergenerates.

[24] See Stalnaker (1999, ch.13 and 14) for further discussion of this response.

[25] See Altrichter (1985) and Sorensen (1996) for further examples of beliefs in metaphysical impossibilities.

4 Rejecting the assumptions

The upshot so far is that given two assumptions—the existence of unknowable propositions and that counterpossibles are vacuously true—OLD overgenerates. This section and the next argue that rejecting either of the two assumptions does not help defenders of OLD. Nothing short of revising OLD will do.

4.1 The first assumption

How might defenders of OLD reject the assumption that there is at least one unknowable proposition? Since the unknowable propositions I discussed were unknowable at least partly due to the entailment from knowledge to truth, they might reject the assumption by rejecting that entailment. Doing this, they say that for every subject x and proposition p, and some metaphysically possible world w, x knows p in w although p is false in w.

Defenders of OLD may suggest that this move is independently motivated. Following Hazlett (2010, p. 501), they may argue that there are utterances whose felicity we can explain only if we reject the claim that knowledge entails truth. Consider

2. Everyone knew that stress caused ulcers, before two Australian doctors in the early 80s proved that ulcers are actually caused by bacterial infection.

Because 2 attributes knowledge of a falsehood to the denotation of 'everyone', the claim that knowledge entails truth would have us predict that it is infelicitous.[26] But it is not. So, the argument goes, the claim has to go.

However, rejecting the claim that knowledge entails truth is costly and Hazlett's motivation for doing so inconclusive. First, the claim that knowledge entails truth explains why utterances like 3 are infelicitous:

3. *Laura knows that it's Monday, but it's not.

In addition, it is unclear whether we can explain these data without saying that knowledge entails truth. We cannot, for instance, explain them by appealing to a conversational implicature to the effect that p is true. Such an implicature could be canceled. But, the infelicity of utterances like 3 shows that it generally cannot be. Rejecting the entailment from knowledge to truth may thus leave us without an explanation of why utterances like 3 are infelicitous.

Even setting aside the cost of rejecting the claim that knowledge entails truth, two responses to the argument against it are available. The first says that, in utterances like 2, occurrences of the verb 'know' denote knowledge* not knowledge. Given this, we can maintain that knowledge, but not knowledge*, entails truth.[27] The second response explains the felicity of utterances like 2 by appeal to protagonist projection. Faced with utterances where the protagonist's perspective involves a false belief—in 2, a belief that stress caused ulcers—we may imagine what seems (or seemed) true from

[26] Strictly speaking, we only get this prediction if we assume that the entailment is common knowledge among competent speakers of English. For ease of exposition, I omit this assumption in the text.

[27] See Turri (2011) and Tsohatzidis (2012).

the protagonist's perspective. We may then treat as felicitous the use of words the protagonist herself would use (or have used) to describe those beliefs—in 2, 'knew'.[28] Thus, we can explain the felicity of utterances like 2, even if knowledge entails truth.

For present purposes I need not settle for either of the responses. Their mere availability shifts the burden of proof onto those who deny that knowledge entails truth. They must explain why the responses fail. As things stand, denying that knowledge entails truth is thus unmotivated. Given this, defenders of OLD should not deny that knowledge entails truth to avoid the prediction that, in suitable contexts, everyone believes all unknowable propositions.

4.2 The second assumption

OLD's prediction that, in suitable contexts, everyone believes all unknowable propositions results, in part, from treating '$\Box\!\!\rightarrow$' as a universal quantifier whose domain is restricted to a subset of metaphysically possible worlds only. Due to this restriction, the domain of '$\Box\!\!\rightarrow$' is empty given a metaphysically impossible antecedent like 'Joan knows a.' This yields 1's vacuous truth and OLD's mistaken predictions.

To block my argument, we may treat '$\Box\!\!\rightarrow$' as a universal quantifier whose domain is restricted to a subset of the metaphysically possible *or impossible* worlds. ('$\phi \Box\!\!\rightarrow \psi$' is now true at a world w just in case all the metaphysically possible or impossible ϕ-worlds closest to w are ψ-worlds.)[29] Given this, and given that there are sufficiently many metaphysically impossible worlds, the domain of '$\Box\!\!\rightarrow$' is non-empty even given a metaphysically impossible antecedent. Thus, counterpossibles are not vacuously true.

This response to my argument may be independently motivated. Some reject the claim that counterpossibles are vacuously true because it leads to strange predictions.[30] Consider

4. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America at the time would have cared.
5. If Hobbes had (secretly) squared the circle, all sick children in the mountains of South America would not have cared.

As it is metaphysically impossible for Hobbes to square the circle, 4 and 5 are counterpossibles. Thus, if counterpossibles are vacuously true, both are true. But intuitively, the sick children at issue would not have known if Hobbes had secretly squared the circle and so could not and would not have cared. So, 4 is false, 5 true.

The revised semantics of would-counterfactuals predicts our intuitions. 4 is now true at our world @ just in case all (impossible) worlds closest to @ at which Hobbes secretly squared the circle are worlds at which sick children in the mountains of South America at the time cared. This sentence is false because in the impossible worlds closest to @ (with respect to the effects of secrecy, say) the sick children at issue did

[28] See Stokke (2013) and Buckwalter (2014).

[29] Berto (2017) offer a semantics along those lines.

[30] Kment (2014) and Berto (2017), for instance. By contrast, Lewis (1973), Emery and Hill (2017), Vetter (2016), and Williamson (2007, 2016), among others, accept the assumption. See Jenny (2018) and Tan (2019) for further arguments in favor of rejecting the assumption.

not know that Hobbes secretly squared the circle and so did not care. By contrast, 5 is now true at @ just in case all (impossible) worlds closest to @ at which Hobbes secretly squared the circle are worlds at which sick children in the mountains of South America at the time did not care. This sentence is true for the reason for which the first was false.

The revised semantics also prevents OLD from predicting that, in suitable contexts, everyone believes all unknowable propositions. That prediction depended on the result that counterfactuals of the form 'If x knew p_u, x would treat p_u in F' (where p_u is a variable ranging over all and only unknowable propositions) are true with an arbitrary way of treating p_u assigned to F. But, by the revised semantics, such counterfactuals are true at a world w just in case all impossible worlds closest to w at which x knows p_u are worlds at which x treats p_u in F. Crucially, this formula is true on some assignments to F, but not others. Assigning Joan to x, a to p_u, and Joan's actual way of treating a to F, for instance, it is false. For in the impossible worlds closest to our world Joan does not (absent devious desires and intentions) know a and deny a when asked whether a, refuse to rely on a in reasoning, and so on. Thus, 1 is false and my argument blocked.

However, I have three concerns about the current response on behalf of OLD. First, Williamson (2016) holds that counterpossibles are vacuously true. He insists that the domain of '$\Box\!\!\rightarrow$' is restricted to metaphysically possible worlds only and offers an error theory to explain the linguistic data meant to motivate the revised semantics. Moreover, Emery and Hill (2017), among others, make similar moves. This highlights that the revised semantics is a significant commitment. Many otherwise friendly to a knowledge-first view may not want to accept it.

Second, some defenders of the revised semantics, e.g. Kment (2014), reject impossible worlds at which logical truths are false. However, absent such worlds, OLD predicts that, in suitable contexts, everyone believes all negations of logical truths. To see this, simply let $\neg l$ be the negation of an arbitrary logical truth and substitute $\neg l$ for a in my argument above. Thus, for OLD not to overgenerate, we must accept not only the revised semantics, but also, contrary to some defenders of that semantics, impossible worlds at which logical truths are false.

The two concerns so far turn on how controversial the current response is. In the next section, I introduce a third, more forceful worry. Even the revised semantics for would-counterfactuals does not keep OLD from making mistaken predictions.

5 A further challenge

We can divide versions of OLD into two camps: relativized and unrelativized. The condition denoted by 'treats p' is relativized to a proposition, p. The conditions denoted by 'behaves' and 'is disposed to act (think, feel)' are unrelativized. As it occurs in OLD, 'Φ' may be substituted with a string denoting a relativized or an unrelativized condition. Thus, we get relativized and unrelativized versions of OLD, which appeal to relativized and unrelativized conditions respectively.

Section 5.1 argues against unrelativized OLD. For concreteness, I consider the version we get by substituting 'behaves' for 'Φ'. Section 5.2 targets relativized versions

of OLD. Though different, my challenges to both turn on would-counterfactuals that are guaranteed to be true due to the entailment from knowledge to truth.[31]

5.1 Unrelativized OLD

Joan suspends judgment about how *exactly* she behaves. She knows that many aspects of her behavior are inaccessible to her and so that she would likely be mistaken if she believed that some fully specific way of behaving is her fully specific way of behaving. Joan, it seems, does not believe that she behaves in her fully specific way of behaving. However, supposing merely that her fully specific way of behaving is contextually salient, unrelativized OLD predicts the opposite.

To see this, let B be Joan's actual fully specific way of behaving. Suppose also B is contextually salient, perhaps because we are asking in what fully specific way Joan behaves. Joan satisfies OLD's first condition. Now consider

6. If Joan knew that she behaves in B, she would behave in B.

That knowledge entails truth guarantees that 6 is true, at least if 6 is evaluated at a metaphysically possible world.[32] On the revised semantics, 6 is true at a world w iff all worlds closest to w at which Joan knows that she behaves in B are worlds at which Joan behaves in B. Supposing w is metaphysically possible, it follows from the entailment from knowledge to truth that at w, for all x and p, x knows p at w only if p is true at w. Because 6 is not a counterpossible—that Joan behaves in B seems to be knowable for Joan—looking for the worlds closest to w at which 6's antecedent is true does not require looking for worlds at which this entailment fails. Plausibly, then, all the worlds closest to w at which 6's antecedent is true are worlds at which it is true that, for all x and p, x knows p only if p is true. Thus, all these worlds are worlds at which Joan behaves in B. So, our revised semantics predicts that 6 is true, at least if evaluated at a metaphysically possible world.

Given this prediction, Joan satisfies OLD's second condition in addition to its first. OLD thus predicts that Joan believes that she behaves in B. Even the revised semantics does not prevent unrelativized OLD from overgenerating.

Moreover, Joan and B are an arbitrary subject and fully specific way of behaving. So, we can generalize. For every subject x and fully specific way of behaving F, if x behaves in F and F is contextually salient, x believes that x behaves in F. Consequently, further counterexamples to unrelativized OLD arise wherever a subject x behaves in a fully specific way F, F is contextually salient, and behaving in F suggests that x suspends judgment about, does not believe, or disbelieves the proposition that x behaves in F.

The problem is compounded because fully specific ways of behaving are cheap. Plausibly, for every (living) subject x, x behaves in some fully specific way.[33] Thus,

[31] Thanks to an anonymous referee for helping me to clarify why these would-counterfactuals are of interest.

[32] Given the standard semantics of would-counterfactuals, this is so because the entailment from knowledge to truth holds at all metaphysically possible worlds and so at all metaphysically possible worlds at which 6's antecedent is true.

[33] Below, I omit the restriction to living subjects.

for every subject x there is some fully specific way of behaving F such that if F is contextually salient, x believes that x behaves in F.

One upshot of this is that, in contexts where the fully specific way of behaving of every subject is salient (perhaps because we ask how exactly each subject behaves), OLD predicts that every subject believes the proposition that results from assigning it and its fully specific way of behaving to 'x behaves in F.' So, whilst adopting our revised semantics of would-counterfactuals avoids the problems of Sect. 3, unrelativized OLD nonetheless predicts that, in suitable contexts, everyone believes some, rather odd, knowable proposition.

It is worth immediately setting aside a familiar response. My argument assumed that propositions which result from assigning a subject and its fully specific way of behaving to 'x behaves in F' are in the domain of the universal quantifier binding unrelativized OLD's proposition variable. So, to block my argument, we might restrict that domain to all and only propositions that do not result from assigning a subject and its fully specific way of behaving to 'x behaves in F.'

However, given this restriction, unrelativized OLD predicts that subjects cannot believe propositions that result from assigning them and their fully specific way of behaving to 'x behaves in F.' Yet it seems that they can. Suppose an omniscient angel reveals herself to Joan and gains her trust by correctly reporting on Joan's occurrent mental episodes. The angel then gives Joan full information about how she behaves. Consequently, Joan comes to believe that she behaves in B. Absent evidence to the contrary, this scenario seems possible. So, it seems that subjects can believe propositions that result from assigning them and their fully specific way of behaving to 'x behaves in F.' Our account of belief, if it is to be fully general, should predict this. Yet on the current response, unrelativized OLD does not and so undergenerates.

5.2 Relativized OLD

I now turn to relativized OLD. For concreteness, I focus on the version we get by substituting 'treats p' for 'Φ'. Yet, as is easily verified, my argument generalizes to other relativized versions of OLD, such as those we obtain by substituting 'behaves with respect to p' or 'is disposed to act (think, feel) with respect to p' for 'Φ.'

Suppose Joan, having reflected on how she treats various propositions about ancient philosophers (including a), adopts a rational policy. She treats the proposition that she treats every proposition in the way in which she treats a ('t') in the way in which she treats a, i.e. D. She denies t, refuses to rely on it in her reasoning, and so on. Joan's adoption of this rational policy strongly suggests that she does not believe t, even that she believes $\neg t$. However, supposing only that D is contextually salient, relativized OLD predicts the opposite.

To see this, suppose D is contextually salient, perhaps because we are asking how Joan treats t. Since Joan treats t in D, she thus meets OLD's first condition. Now consider:

7. If Joan knew t (the proposition that she treats every proposition in D), she would treat t in D.[34]

7 is intuitively true, given the entailment from knowledge to truth (and an elimination rule for the universal quantifier). So, Joan satisfies OLD's second condition. OLD therefore predicts that Joan believes t.

Does our revised semantics predict that 7 is true? To answer this question, notice first that 7 is a counterpossible because t is unknowable for Joan. To see that it is, suppose for reductio Joan can know t. Then, Joan knows t at some metaphysically possible world—call it m. By the entailment from knowledge to truth, Joan treats every proposition in way D at m. But now let p be an arbitrary proposition. Joan treats p in D at m, $\neg p$ in D at m, $p \wedge \neg p$ in D at m, and so on. Joan treats all propositions alike at m, whether they are contraries, contradictories or whatever.

Joan's psychology at m is strange. Intuitively, it is inconsistent with Joan knowing any proposition at m. For knowing p seems to entail, at all possible worlds, that one treats p and $\neg p$, for instance, differently. In some sense, one's knowing p rules out $\neg p$ for one. (Perhaps by setting the evidential probability of p at 1, that of $\neg p$ at 0.) And that must be reflected, somehow, in how one treats p and $\neg p$. Yet, in Joan's psychology at m, it is not. Thus, Joan does not know t at m.

Since our initial supposition entailed that Joan does know t at m, we now have a contradiction and something has to give. Holding fixed the entailment from knowledge to truth and the intuition about Joan's psychology, we must say that t is unknowable for Joan.

As counterpossibles are not vacuously true on the revised semantics, defenders of relativized OLD may hope that the revised semantics predicts that 7 is false. (Although that outcome may worry defenders of the revised semantics.) However, given a plausible assumption, the revised semantics predicts that 7 is true.

On the revised semantics, 7 is true at a world w just in case all worlds closest to w at which Joan knows t are worlds at which Joan treats t in D. Since 7 is a counterpossible, looking for the closest worlds at which its antecedent is true requires looking for metaphysically impossible worlds. Thus, we must give up something true at all metaphysically possible worlds to find the closest worlds at which 7's antecedent is true.

There are two candidates. First, the entailment from knowledge to truth. Second, the intuitive datum about Joan's psychology. How our revised semantics evaluates 7 at w depends on which of the two we give up. If we give up the entailment from knowledge to truth, 7 is false. To find the closest worlds at which 7's antecedent is true, we now look for the worlds closest to w at which it is false that for all x and p, x knows p only if p is true. (Because these worlds are closest to w and at w the intuitive datum about Joan's psychology holds, the datum holds at these worlds.) Consequently, at some of the closest worlds at which 7's antecedent is true, its consequent is false.

By contrast, if we give up the intuitive datum about Joan's psychology, 7 is true. By the entailment from knowledge to truth, it is true at w that for all x and p, x knows p at w only if p is true at w. To find the worlds closest to w at which 7's antecedent is

[34] On the intended reading, 'every' does not receive an implicit domain restriction—its domain thus contains all and only propositions, not some proper subset of them.

true, we now look for worlds at which the intuition about Joan's psychology is false. (Because those worlds are closest to w and at w the entailment from knowledge to truth holds, the entailment holds at those worlds.) Consequently, all closest worlds at which 7's antecedent is true are worlds at which its consequent is true.

Which of the two candidates do we give up? 7 is, intuitively, true. So, when looking for the closest worlds at which 7's antecedent is true, we intuitively give up the intuition about Joan's psychology, not the entailment from knowledge to truth. This is perhaps not unexpected. For it is a platitude that "in knowledge, mind is adapted to world" (Williamson 2000, p. 1). This platitude suggests that the connection between knowing p and p's truth is more entrenched in (our conception of) knowing p than the connection between knowing p and how one treats p, $\neg p$, and so on.

Taking stock, given that we give up the intuitive datum about Joan's psychology in looking for the worlds relevant to evaluating 7, the revised semantics of would-counterfactuals predicts that 7 is true. Yet given 7's truth and that D is contextually salient, OLD mistakenly predicts that Joan believes t.

Moreover, Joan and D were selected arbitrarily amongst subjects and ways of treating a proposition. So, we can generalize. For every subject x and way of treating a proposition F, if F is contextually salient and x treats the proposition that she treats every proposition in F in F, x believes that she treats every proposition in F. Further counterexamples to relativized OLD thus arise wherever a subject x treats the proposition that x treats every proposition in F in F, F is contextually salient, and treating a proposition in F suggests that x suspends judgment about, does not believe, or disbelieves that proposition.

The problem is compounded because treating the proposition that one treats every proposition in D in D is a rational policy. So, plausibly, for every rational subject x_R, x_R treats the proposition that x_R treats every proposition in D in D. But now, it follows that for every rational subject x_R, if D is contextually salient, x_R believes that x_R treats every proposition in D. Since this proposition is unknowable for x_R, relativized OLD predicts that, in suitable contexts, every rational subject believes some, rather odd, unknowable proposition.

It is worth setting aside a, by now very familiar, response. I assumed that propositions that result from assigning a subject and way of treating a proposition to 'x treats every proposition in F' are in the domain of the universal quantifier binding OLD's proposition variable. So, to block my argument, we might restrict that domain to all and only propositions that do not result from assigning a subject and way of treating a proposition to 'x treats every proposition in F.'

However, given this restriction, OLD predicts that subjects cannot believe propositions that result from assigning a subject and way of treating a proposition to 'x treats every proposition in F.' But they can. To see this, suppose Joseph is an extreme contrarian and treats every p by asserting $\neg p$ whenever someone utters p. Neither Joseph nor his colleagues are under any illusion about this: they truly believe that he treats every p by asserting $\neg p$ whenever someone utters p. Absent evidence to the contrary, this scenario seems possible. So, it seems that, more generally, one can believe propositions that result from assigning a subject and way of treating a proposition to 'x treats every proposition in F.' Our account of belief, if it is to be fully general, should predict this. But, given the current response, OLD does not and so undergenerates.

6 A better alternative

The story so far is this. OLD overgenerates, given two orthodox assumptions. Rejecting the first assumption is unmotivated, rejecting the second insufficient to keep OLD out of trouble. Nothing short of revising OLD seems to do. This section shows that NEW avoids OLD's difficulties and thus is preferable to OLD.

Several responses on behalf of OLD restricted the domain of the universal quantifier binding OLD's proposition variable—call them the restriction responses. The restriction responses led OLD to undergenerate. As OLD appeals to just one proposition variable, bound by a universal quantifier, only one's satisfaction of a counterfactual involving knowledge of p can underwrite a prediction that one believes p. If the domain of the universal quantifier is then restricted to a proper subset of the set of propositions, the propositions that are not members of that subset fall outside the account's purview. One's satisfaction of a counterfactual involving knowledge of a proposition not in the universal quantifier's domain cannot underwrite a prediction that one believes that proposition.

To avoid this concern for the restriction responses, we need not abandon them however. An alternative is to supplement them with further claims. This is just what NEW, reproduced here, does.

NEW For all x and p, x believes p iff: for some knowable q and contextually salient way F, (i) x Φs with respect to p in F and (ii) if x knew q, x would Φ with respect to q in F.

NEW is strict as the domain of the existential quantifier binding q is restricted by the predicate 'knowable.' Consequently, one's satisfaction of a would-counterfactual involving knowledge of unknowable propositions cannot underwrite a prediction that one believes p. NEW's strictness is a built-in restriction response. It prevents NEW both from predicting that, in suitable contexts, everyone believes all unknowable propositions, even without the revised semantics of would-counterfactuals, and from predicting that, in suitable contexts, every rational subject believes some unknowable proposition.

Yet NEW's built-in restriction response does not limit its purview because NEW is flexible. First, NEW does not exclude any propositions from the domain of the universal quantifier binding p. Second, due to NEW's appeal to two proposition variables, p bound by a universal, q bound by a restricted existential quantifier, one's satisfaction of a counterfactual involving knowledge of q can underwrite a prediction that one believes p (where, if p is unknowable, q is distinct from p, else may be identical to p). As a result, one's satisfaction of a counterfactual involving knowledge of a knowable proposition can underwrite a prediction that one believes an unknowable proposition. So, NEW does not predict that one cannot believe unknowable propositions.

Of course, NEW's strictness and flexibility do not speak to my challenge for unrelativized OLD involving knowable propositions in Sect. 5.1. To deal with this challenge, NEW is relativized insofar as the condition denoted by 'Φ with respect to p' is. Because of this, Joan's satisfaction of 6, 'If Joan knew that she behaves in B, she would behave in B' cannot underwrite a prediction that she believes that she behaves in B. This prevents NEW from predicting that, in suitable contexts, every subject believes the

proposition that results from assigning it and its fully specific way of behaving to 'x behaves in F.'

At the same time, however, NEW does not predict that one cannot believe the proposition that results from assigning one and one's fully specific way of behaving to 'x behaves in F.' Joan, for instance, can believe that she behaves in B so long as there is a contextually salient way F such that she Φs with respect to the proposition that she behaves in B in F and if she knew that proposition, she would Φ with respect to that proposition in F.

In sum, NEW avoids under- and overgenerating in the way OLD did. Thus, NEW is preferable to OLD. Whether we should, all things considered, accept it, however, remains an open question. To answer it, we must substitute for 'Φ', interpret 'Φing *with respect to* p', and do much else. These are tasks I leave for another occasion.

Acknowledgements For reading several earlier drafts, I am grateful to Guy Longworth. Thanks also to Mark Jago, Tristan Kreetz, Jennifer Nagel, Johannes Roessler, David Strohmaier, Tim Williamson, and three anonymous referees for this journal for helpful comments, audiences at Nottingham, Oxford, Pavia, Porto and Warwick for discussion, and Jack Blaicklock for proofreading.

References

Altrichter, F. (1985). Belief and possibility. *The Journal of Philosophy*, *82*(7), 364.
Barker-Plummer, D et al. (2011). Language, proof, and logic. eng. 2. ed. OCLC: 726620713. Stanford, Calif: CSLI Publ. ISBN: 978-1-57586-632-1.
Berto, F., et al. (2017). Williamson on counterpossibles. *Journal of Philosophical Logic*, *0022–3611*, 1573.
Braithwaite, R. B. (1932). The nature of believing. *Proceedings of the Aristotelian Society*, *33*, 129–146.
Buckwalter, W. (2014). Factive verbs and protagonist projection. *Episteme*, *11*(4), 391–409.
Butterfill, S. (2013). What does knowledge explain? Commentary on Jennifer Nagel 'Knowledge as a Mental State'. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology 4* (pp. 309–320). Oxford: Oxford University Press.
Bcking, S. (2017). Composing wie wennthe semantics of hypothetical comparison clauses in German. *Natural Language & Linguistic Theory*, *35*(4), 979–1025.
Correia, F. (2017). Real definitions. *Philosophical Issues*, *27*(1), 52–73.
Dorr, C. (2016). To Be F Is To Be G. *Philosophical Perspectives*, *30*(1), 39–134.
Dunn, R., & Suter, G. (1977). Zeno vendler on the objects of knowledge and belief. *Canadian Journal of Philosophy*, *7*(1), 103–114.
Emery, B. N., & Hill, C. S. (2017). Impossible worlds and metaphysical explanation: Comments on kments modality and explanatory reasoning. *Analysis*, *77*(1), 134–148.
Geurts, B (2008). Existential import. In Existence: Semantics and syntax. Studies in Linguistics and Philosophy (pp. 253–271). Dordrecht: Springer. ISBN: 978-1-4020-6198-1 978-1-4020-6197-4.
Greenough, P., & Pritchard, D. (Eds.). (2009). *Williamson on knowledge*. Oxford: Oxford University Press.
Hawthorne, J., Rothschild, D., & Spectre, L. (2016). Belief is weak. *Philosophical Studies*, *173*(5), 1393–1404.
Hazlett, A. (2010). The myth of factive verbs. *Philosophy and Phenomenological Research*, *80*(3), 497–522.
Holton, R. (2017). IFacts, factives, and contrafactives. *Aristotelian Society Supplementary*, *91*(1), 245–266.
Hyman, J. (2015). *Action, knowledge, and will*. Oxford: Oxford University Press.
Hyman, J. (2017). IIKnowledge and belief. *Aristotelian Society Supplementary*, *91*(1), 267–288.
Jenny, M. (2018). Counterpossibles in science: The case of relative computability. *Nos*, *52*(3), 530–560.

Jones, O. R. (1975). Can one believe what one knows? *Philosophical Review*, *84*(2), 220–235.
Kment, B. (2014). *Modality and explanatory reasoning*. Oxford: Oxford University Press.
Kratzer, A. (1979). Conditional necessity and possibility. In *Semantics from different points of view* (pp. 117–147). Dordrecht: Springer.
Lewis, D. K. (1973). *Counterfactuals*. London: Blackwell.
Marcus, R. B. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research*, *50*, 133.
McGlynn, A. (2017). Mindreading knowledge. In J. Adam Carter, E. C. Gordon, & B. Jarvis (Eds.), *Knowledge first: Approaches in epistemology and mind* (pp. 72–94). New York: Oxford University Press. ISBN: 978-0-19-871631-0.
Nagel, J. (2013). Knowledge as a mental state. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology 4* (pp. 273–308). Oxford: Oxford University Press.
Nagel, J. (2017). Factive and nonfactive mental state attribution. *Mind & Language*, *32*(5), 525–544.
Roessler, J. (2013). Knowledge, causal explanation, and teleology. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology 4* (pp. 321–333). Oxford: Oxford University Press.
Rose, D. (2015). Belief is prior to knowledge. *Episteme*, *12*(3), 385–399.
Rosen, G. (2015). Real definition. *Analytic Philosophy*, *56*(3), 189–209.
Rosenthal, D. M. (1976). Res cogitans: An essay in rational psychology. *Journal of Philosophy*, *73*(9), 240–252.
Rysiew, P. (2013). Is knowledge a non-composite mental state? In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology 4* (pp. 334–344). Oxford: Oxford University Press.
Sorensen, R. A. (1996). Modal bloopers: Why believable impossibilities are necessary. *American Philosophical Quarterly*, *33*(3), 247–261.
Stalnaker, R. (1999). *Context and content: Essays on intentionality in speech and thought. Oxford cognitive science series*. Oxford: Oxford University Press. ISBN: 978-0-19-823708-2 978-0-19-823707-5.
Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Blackwell.
Stalnaker, R. C. (1987). *Inquiry. eng. 1. MIT Press paperback ed. A Bradford book. OCLC: 833263071*. Cambridge: MIT Press.
Stokke, A. (2013). Protagonist projection. *Mind & Language*, *28*(2), 204–232.
Tan, P. (2019). Counterpossible non-vacuity in scientific practice. *The Journal of Philosophy*, *116*(1), 32–60.
Tsohatzidis, S. L. (2012). How to forget that 'Know' is factive. *Acta Analytica*, *27*(4), 449–459.
Turri, J. (2011). Mythology of the factive. *Logos and Episteme*, *2*(1), 143–152.
Umbach, C., & Gust, H. (2014). Similarity demonstratives. *Lingua. SI: Modification at the Interfaces*, *149*, 74–93.
Vendler, Z. (1972). *Res cogitans: An essay in rational psychology. Contemporary philosophy*. Ithaca: Cornell University Press. ISBN: 978-0-8014-0743-7.
Vetter, B. (2016). Counterpossibles (not only) for dispositionalists. *Philosophical Studies*, *173*(10), 2681–2700.
Williamson, T (1996). Vagueness. Problems of philosophy. New York: Routledge. ISBN: 978-0-415-03331-2 978-0-415-13980-9
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell Publishing Ltd.
Williamson, T. (2016). Counterpossibles. In Topoi, pp. 1–12. issn: 1572-8749.