Self-Knowledge of Belief Requires Understanding of Propositions

Lukas Schwengerer

Penultimate Draft – Please cite the published version at

https://link.springer.com/article/10.1007/s10670-024-00917-1

Abstract

I show that from common views about propositions as sets of possible worlds and knowledge requiring a sufficiently strong safety condition one can derive a condition stating that self-knowledge of belief is only possible if the content of that belief is fully understood. I show this by a reductio. If a subject S lacks full understanding of a proposition *p*, then S's belief about believing that p cannot amount to knowledge. Even though my argument is based on particular views about propositions and knowledge, I argue that the same kind of argument is also available for other knowledge conditions that rule out relevant luck and other accounts of propositions. However, for many of these other accounts of knowledge the requirement for understanding *p* will not be full understanding, but only sufficiently high understanding. These results tell us how self-knowledge, ruling out luck and understanding belief contents relate.

Introduction

I show that one can derive a condition for self-knowledge of belief from common views about propositions and knowledge. In particular, if we understand propositions as sets of possible worlds with Stalnaker (1984; 1999), and accept a fairly strong safety condition for knowledge with Sosa (1999) or Pritchard (2005; 2007; 2012), then knowing that I believe that p requires that I fully understand the proposition *p*. I show this by a reductio. If a subject S lacks full understanding of a proposition *p*, then S's belief about believing that *p* cannot amount to

knowledge. Even though my argument is put forward within the frameworks of Stalnaker and strong safety, I argue that the same kind of argument is applicable to other knowledge conditions that rule out relevant luck and other accounts of propositions, although with slightly different results. I show this by considering weaker safety accounts and Sosa's (2007; 2010; 2015) virtue reliabilism.

The Argument

Let me start with the general form of the argument, before going through the individual steps. Within the argument I indicate mere assumptions without any significant theoretical baggage with 'N_A'. I indicate more heavyweight premises with theoretical commitments with 'N_P' and conclusions derived within the argument with 'N_C'.

1_A) Subject S has a belief-forming process that generates a belief that S believes that *p*.[1]

2_P) A fairly strong safety condition for knowledge: If S knows that *p* then S's true belief that *p* could not have easily been false. More precisely, in all close possible worlds (or at least in all very close possible worlds) in which S believes that *p* via the same method of belief formation M that S uses in the actual world, *p* is true.[2]

[1] This assumption is relatively neutral on accounts of self-belief, but not completely neutral. It is incompatible with some accounts of self-knowledge that suggest that first-order beliefs are identical with or already logically entail beliefs about the respective first-order beliefs. So, my belief that p would at the same time be, or entail a belief that I believe that p. One such view is proposed by Boyle (2011).

[2] Safety can be formulated in different ways and not every way of spelling out the safety condition will work here. The strength of a safety condition is captured by the number of close possible worlds in which S believes that *p* (by the same method as in the actual world) and *p* is false, while the belief that *p* still qualifies as safe.
A very strong safety condition requires that in *all* close possible worlds in which S believes *p* (by the same method as in the actual world) *p* is true. Weaker conditions require only that *p* is true in *most* of these close possible worlds. The more of these close possible worlds in which *p* is false are tolerated without failing safety, the weaker the safety condition. The strength of a safety condition also depends on the range of possible worlds considered as relevant for safety. The strongest safety condition takes all possible worlds to be relevant, weaker version only close possible worlds (with lots of room for how to exactly draw the borders for closeness).
Sufficiently strong versions of safety for my purpose are, for instance, proposed by Sosa (1999), Grundmann (2020) or Pritchard (2012). Sosa (1999) presents the strongest form, Pritchard a weaker one that only requires *p* to be true in all very close possible worlds and in most close possible worlds for the belief that *p* to be safe. In

3_P) A modal understanding of propositions: A proposition is understood as a set of possible worlds in which the proposition is true.

4_P) For any non-trivial proposition *p* there is a very similar proposition *p**, such that without full understanding of p S would not be able to tell *p* and *p** apart.

5_A) Suppose that S lacks full understanding of *p*.

6_P) If a subject S cannot tell *p* and *p** apart due to a lack of understanding, then if S believes *p* in *w* there must be a close world *w'* where S believes *p** instead of *p*.

7_C) S's belief that S believes that *p* could have easily been false, because S could have been believing that *p** instead of believing *p* and would not have noticed the difference. (from 4_P, 5_A, 6_P)

8_C) S's belief that S believes that *p* is not safe, because it could have easily been false. (from 2_P, 7_C)

9_C) S does not know that S believes that *p*, but merely believes that S believes that *p*. (from 1_A, 2_P, 8_C)

(1_A) is merely the assumption that one can form beliefs about one's own beliefs. Premise (2_P) is a simple version of a strong safety condition. Premise (3_P) is Stalnaker's conception of a

---

my argument I use a relatively strong version ('in *all* close possible worlds'). I will come back to weaker versions of safety.

My formulation of choice for safety also refers to the method involved. Method safety theories can distinguish between local and global safety. I take Bernecker's (2020) formulation to illustrate this difference:

> Local Reliability of Methods: Method M by which S forms (sustains) the belief that p is locally reliable if and only if in all nearby worlds where S employs M it yields the belief that p only if p is true.
>
> Global Reliability of Methods: Method M by which S forms (sustains) the belief that p is globally reliable if and only if in all nearby worlds where S employs M it yields the belief that p only if p is true and it yields only true beliefs in a range of propositions relevantly similar to p. (Bernecker, 2020, p. 5104)

I do not think it matters much for my argument, but I prefer global safety over local. Hence, one may add a clause like 'and it yields only true beliefs in a range of propositions relevantly similar to *p*' to my formulation of safety. Global safety comes with the advantage that a method that safely generates beliefs about *p* will also get things right about the very similar proposition *p**.

proposition. These are supplemented with two rather uncontroversial premises in (4_P) and (5_A) and the slightly more controversial (6_P). Premise (4_P) states that it is possible for two non-trivial propositions to exist without an agent being able to tell them apart if that agent lacks some understanding of *p*. In Stalnaker's framework we get this plausibly as follows. First take a non-trivial proposition *p* that is not fully understood by S. Not fully understanding entails that for some possible world S will not be in a position to know whether *p* is true or false. This is indicated by Stalnaker's discussion of degrees of understanding propositions (Stalnaker, 1984, p. 65). Take this possible world that S lacks sufficient epistemic access to and change the truth value provided by *p* for this one world. With this move we arrive at a very similar but nevertheless different proposition *p**, which has the same truth value as *p* in all possible worlds except one. The subject S cannot tell *p* and *p** apart, because the agent does not know and is in no position to know the truth value of *p* for the one world in which *p* and *p** differ. Both propositions look the same to the agent, even though they determine different sets of possible worlds in which they are true. The agent would react the same to an assertion that *p* and to an assertion that *p**. For any proposition that an agent does not fully understand we can create such a proposition *p**, because a lack of understanding of *p* in this framework is equal to not being in a position to know the truth value for *p* in some possible world. The qualifier non-trivial is meant to rule out propositions for which any change to the truth value in any possible world would always be noticeable, such as necessary truths.[3] Premise (5_A) now stipulates that the subject S is exactly in a case as proposed in (4_P). S lacks full understanding of *p* and hence could not tell *p* and *p** apart.

[3] These are tricky for Stalnaker's view in general. I will bracket this issue as it is not important for my argument.

To get from (4_P) and (5_A) to (7_C) a little more work is needed. (7_C) requires that S's belief that S believes *p* could have easily been false. In other words, it requires a close possible world in which S believes that S believes *p*, but S actually believes *p**. (4_P) and (5_A) entail that in some possible world S falsely believes that S believes *p*, but they do not establish that it is a close possible world. (6_P) bridges the gap to (7_C) by establishing the closeness of the possible world in question. (6_P) is at least not obvious, so why should one accept it? The best reason for (6_P) comes from thinking about modal closeness. Modal closeness is usually – following Lewis (1973) – understood in terms of world similarity. But world similarity itself is notoriously vague. Lewis himself argues that this vagueness cannot and should not be completely removed (1973, pp. 91-95). However, the vagueness can be limited such that the modal closeness required in (6_P) becomes apparent. Lewis presents the idea behind his understanding of counterfactuals at the very beginning of *Counterfactuals*. He writes:

> 'If kangaroos had no tails, they would topple over' seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tail, and which resembles our actual state of affairs as much as kangaroos having no tail permits it to, the kangaroos topple over. (Lewis, 1973, p. 1)

The important part for my purpose is the clause referring to the resemblance to our actual state of affairs to the degree that kangaroos having no tail permits. This points to the world similarity in question. When one thinks of a counterfactual situation, one holds as much of the actual world fixed as possible. The more one can hold fixed, the more similarity that possible world has to the actual world. Modal closeness is therefore based on a question: how much needs to change for the counterfactual to be true? The less changes required, the closer a possible world. This does not provide an exact measure for what counts as a close possible world, but it establishes the tool with which we can make reasonable judgments about modal

closeness. With this in mind I can now show why (6_P) should be accepted. Start with the actual world in which S lacks full understanding of *p* and *p** and therefore cannot tell them apart. Suppose that S believes that *p* in the actual world. How much change in the world is required for a possible world in which S believes *p** instead? How much similarity is permitted by S believing *p**? It seems to me that almost no change is required outside of S's belief that *p**. Given that S is unable to tell the difference, even S's behaviour remains the same. Which minor changes are required exactly might depend on the proposition in question. In a common case S's experiences used to learn the concepts involved in the belief that *p* underdetermine an answer to the question whether *p* or *p** obtains, but S formed the belief that *p* anyway. In such a case tiny changes in the first-order belief formation will be enough for a world in which S believes *p**, which makes the world in which S believes *p** very close to the actual world.[4] With (6_P) in place (7_C) – (9_C) follow from the premises (1_A) – (6_P). Without full understanding of *p*, S's belief that S believes that *p* could easily be false, and hence is not safe. Without safety no knowledge, so S lacks self-knowledge of S's belief that *p*.

Let me illustrate the general idea with a concrete example. Consider the proposition with the content 'Cameron has arthritis.' Suppose I do not fully understand this proposition because I do not fully understand the concept 'arthritis'. I know some common symptoms and I can use the term correctly most of the time, but I do not know what exactly occurs in a body that has arthritis. So, for many instances I would be able to tell whether someone has arthritis, but not for all instances. Sometimes I would not know whether a person has arthritis or does not have

[4] These changes do not impact the method used to form the second-order belief about the first-order belief. Hence, it is not a problem for holding the method generating the second-order belief fixed when considering possible worlds in which S believes that S believes *p*, but S actually believes *p**.

it. Now further suppose I believe that Cameron has arthritis and I want to find out whether I believe that I believe Cameron has arthritis.

Because I partially understand 'arthritis' I am in a position to know whether 'Cameron has arthritis' is true for most possible worlds, but not for all possible worlds. There is at least one possible world in which I would not be able to tell. Let W_2 be a world in which I would not be able to tell and assume that in W_2 Cameron does in fact have arthritis.

Now we can build a second, slightly different proposition – the *p** of the general argument – by stipulating that the new proposition is false in W_2. The arthritis-proposition gives us a set of possible worlds in which it is true, and we now build a new set that includes all the same possible worlds except W_2. This new set that we get by just changing the old set with respect to one world determines a new proposition. Let us say it is the proposition with the content 'Cameron has smarthritis'. I cannot tell this new proposition apart from the old 'Cameron has arthritis' one, because based on my understanding they look exactly the same in all possible worlds in which I would be able to tell whether the proposition is true. If this is right, then my belief that I believe Cameron has arthritis could have been easily false. I could have in fact believed that Cameron has smarthritis and I would still have formed the belief that I believe that he has arthritis. My belief-forming process is unable to detect the difference. If I correctly form the belief that I believe Cameron has arthritis, it will be partially by luck. Hence, I cannot know that I believe Cameron has arthritis.

The example illustrates the general case. In the general version the subject S and the propositions *p* and *p** are completely arbitrary and therefore the argument generalises for all subjects and all non-trivial propositions. Any of S's beliefs about S's own non-trivial beliefs fails

to be safe if S does not understand the proposition in the first-order belief fully. Hence, I propose the following:

> **FULL UNDERSTANDING**: If the modal conception of propositions and a sufficiently strong safety condition for knowledge are true, then for any non-trivial belief that *p* a subject S cannot know that S believes that *p* without fully understanding *p*.

Weak Safety Accounts

In spelling out (2_P) I used a strong version of the safety condition. That is, a condition that requires the belief in all close (or at least all very close) possible worlds to be true. This is not the only version of safety available. Weaker versions only hold that the belief has to be true in most close possible worlds (e.g. Mortini (2022)). The weaker safety condition is not bothered by a single close possible world in which S believes that S believes *p*, but S actually believes *p**. S's belief that S believes *p* is still true in most close possible worlds, even if it is false in one such world.

Have I just provided an argument for weak safety over strong safety? Perhaps, but the advantage of weak safety is rather small. While weaker safety accounts escape my original argument, they do not escape a very similar argument for a requirement of a threshold for understanding. The strategy for that argument is much like my original strategy to generate *p**. A subject with almost full understanding might only lack the ability to distinguish between *p* and *p**. A subject with a slightly bigger gap in their understanding might in addition also lack the ability to distinguish between *p* and *p***, where *p*** is a proposition that is determined by the set of all possible worlds in which *p* is true, except that it differs in one possible world – but a world in which *p* and *p** have the same truth value. For instance, *p* and *p** might differ in W_2, but *p* and *p*** differ in W_3. In the concrete example I now cannot distinguish between

Cameron having arthritis, smarthritis, and omarthritis! The bigger the gap in S's understanding, the more such different but indistinguishable-to-S propositions. At some point this gap is large enough such that in too many close possible worlds in which S believes that S believes *p*, this second-order belief is false. In one close world S actually believes *p**, in another *p***, and so on. To satisfy a weaker safety condition for S's belief that S believes *p*, S needs sufficient understanding of *p*. It might not be full understanding for weaker safety accounts, but nevertheless sufficiently high understanding. How much understanding is required depends on the threshold for 'most close possible worlds.' Usually weak safety accounts do not spell out their notion of 'most close possible worlds' in detail. However, regardless of what counts as most possible worlds, there will always be a degree of understanding a proposition that is required in order to know one's own belief that *p*. The weaker the safety condition for knowledge, the lower the threshold for understanding the proposition. A weaker condition is compatible with more close possible worlds in which S has a false belief before the safety condition fails to be satisfied. Hence, the number of propositions that S cannot distinguish needs to be higher for an unsafe belief. On the other hand, when the demands on what counts as 'most close possible worlds' are high the requirement on understanding *p* will be very close to full understanding. Hence, I propose the following principle:

> **UNDERSTANDING THRESHOLD**: If the modal conception of propositions and a weak safety condition for knowledge are true, then for any non-trivial belief that *p* a subject S cannot know that S believes that *p* without surpassing a threshold for understanding *p*.

As the discussion of weak safety shows, on some accounts we need not require full understanding but merely sufficient understanding. UNDERSTANDING THRESHOLD captures this idea. The understanding required cannot be too low, otherwise it the connection between

knowledge and truth becomes too weak and the problematic form of epistemic luck is not ruled out anymore.

UNDERSTANDING THRESHOLD is also compatible with contextualism about the strength of the relevant safety condition. In general, contextualists hold that knowledge ascriptions depend on the context of the attributor who makes a knowledge ascription. Most prominently, some contextualists argue that the practical stake related to the truth of p matters for attributing knowledge that *p* to a subject. However, other factors of the context can be relevant as well, such as considerations of the salience of error possibilities. I will limit myself to the practical stakes to illustrate the contextualist position.

The basic idea is that the claim 'S knows that *p*' can be true or false depending on how much is at stake for S in the attributor's view. Suppose I attribute to myself that I know the bank will be open tomorrow. If I take the stakes to be very low – if the bank being closed would not matter much anyway – then it takes little for my self-ascription of knowledge to be true. However, if the stakes are high – if the bank being closed tomorrow would be a catastrophe for me – then ascribing knowledge to myself would be quite demanding. I would need especially good evidence that ensures me that the bank will be open tomorrow. The attributor's perception of the relevant stakes is part of the context that determines the truth of the claim. In this case, I am self-attributing, but the general picture works the same with an attributor that differs from the attributee. As DeRose explains the position, "[…] the truth conditions of sentences of the form 'S knows that p' or 'S does not know that p' vary in certain ways according to the context in which the sentences are uttered" (DeRose, 1992, p. 914). When the attributor takes the stakes to be high then knowledge ascriptions become very demanding.

Contextualists can accept versions of safety that are context-sensitive in a way that captures the higher demands in particular contexts. DeRose (2017) holds that the standard for safety shifts with the context. In high-stakes contexts a belief that p needs to satisfy a strong version of safety compared to low-stakes contexts. Strong versions of safety only tolerate very few (if any) close possible worlds in which the subject believes that *p* (via the same method as in the actual world) and *p* is false.[5] The strongest versions of safety tolerate no such close possible worlds. Weaker versions of safety require merely that S's belief that p is true in most close possible worlds in which S believes that p (via the same method as in the actual world). Hence, weak safety tolerates some close possible worlds in which S believes p (via the same method as in the actual world) and p is false. The amount of such 'error worlds' that are tolerated determines the strength of a safety condition. Contextualist safety now suggests that the strength of the safety condition varies with the context of anyone attributing knowledge. If I take the stakes for S to be very high and I attribute knowledge that p to S, then my attribution will only be true if S's belief that *p* is strongly safe. That is, my knowledge attribution is only true if S's belief is true in (almost) all close possible worlds in which S believes that p via the same method as in the actual world. On the other hand, if I take the stakes to be low, then a weaker safety is sufficient for S's belief. I can correctly attribute knowledge even if S's belief could be comparatively easily false.

This idea is captured by the following safety conception of epistemic contextualism taken from Blome-Tillman (2020):

DeRose's Safety Conception of EC

[5] See also footnote 2.

If x satisfies 'knows p' in C, then x's belief that p is safe enough to count as satisfying 'knows p' in C.

The context C determines what counts as safe enough for knowledge. For my argument, that means that the amount of understanding of *p* required to be able to satisfy 'knows that one believes that *p'* is also determined by context C. If the stakes are sufficiently low, then weak safety is enough, such that failing to discriminate between *p*, *p**, *p***,… does not threaten safety and self-knowledge anymore.

A Note on Partial Understanding and an Exception to the Argument

Before generalizing the results let me come back to the role of partial understanding in the argument. I introduced the subject's lack of full understanding as an assumption with (5_A) and I work with Stalnaker's conception of degrees of understanding (Stalnaker, 1984, p. 65). In this conception lacking full understanding entails that for some possible world S will not be in a position to know whether *p* is true or false. Importantly, this does not need to stop a belief to have a particular content. One can hold beliefs without full understanding. This is important because if beliefs themselves would already require full understanding, the question of self-knowledge of partially understood beliefs would be a non-starter.

The idea that propositional attitudes can be held without full understanding is commonly associated with Tyler Burge's work on the topic – and my example of Cameron and arthritis or smarthritis is certainly inspired by Burge's (1979) own arthritis example. In his *Individualism and the Mental* (Burge, 1979) he convincingly argues for content externalism in relation to propositional attitudes. That is, the content of one's attitude can be (and usually is) determined by causal relations to the environment and one can have such an attitude without knowing of these causal relations that enable the attitude. Importantly, one can have such an

attitude even if one partially misunderstands[6] the concepts involved. One can have a belief about Arthritis, even if one partially misunderstands what Arthritis is. Within that discussion, Burge himself is somewhat vague on what understanding exactly means, writing that "[…]'understanding (mastering) a notion' is to be construed more or less intuitively" (Burge, 1979, p. 75), but Burge's notion seems to be compatible with Stalnaker's proposal of modelling understanding on being able to discriminate whether *p* for possible worlds. This model captures an understanding of what something is and/or under what condition something is the case. And Burge is clear that one can have an attitude that *p* without fully understanding the content. I can have attitudes about 'mortgage' without full understanding of what a mortgage is. I just need enough understanding to competently use it in enough ordinary circumstances. Mastery is not required.

While Burge is a clear ally with regard to agents having attitudes without full understanding, Burge might look like an opponent for my argument on a different axis. In *Individualism and Self-Knowledge* (Burge, 1988) argues – among other things – for the possibility of self-knowledge of thoughts even though one only partially understands their contents. At first sight, this looks to be exactly the opposite of what I have been arguing for. However, things are not as bad as they initially appear. The self-knowledge Burge focuses on is slightly different from the self-knowledge I am interested in. Burge is dealing with Descartes-inspired cogito-like judgments.[7] Cogito-like judgments are paradigmatic forms of authoritative and at least partially non-empirical self-ascriptions that are in an interesting way self-referential and self-

[6] Burge holds that the same considerations work for partial understanding that does not include misunderstanding. "Partial understanding is as good as misunderstanding for our purposes" (Burge, 1979, p. 83).
[7] Burge calls self-knowledge of cogito-like judgment 'basic self-knowledge' (Burge, 1988, p. 649), which I take to be rather misleading. They seem more like a special case of self-knowledge to me.

verifying. Examples of such cogito-like judgments are the following (taken from Borgoni (2018)):

> I am now thinking.
>
> I [hereby] judge that Los Angeles is at the same latitude as North Africa.
>
> I [hereby] intend to go to the opera tonight.

Cogito-like judgments are mental self-ascriptions of a particular kind. In performing them one ascribes a mental occurrence and that ascription is made true in the performance of the ascription itself. My judgment that I am now thinking itself makes it true that I am now thinking. Similarly, my judgment that I judge that Los Angeles is at the same latitude as North Africa makes it the case that I judge that. This is different from self-ascribing standing states, such as standing beliefs. Take the following example:

> I believe that Rome is in Italy.

This self-ascription is not cogito-like. It does not make itself true merely in virtue of performing the self-ascription.[8] This difference is important. Burge (1988) argues that cogito-like judgments lead to self-knowledge even when one only partially understands the content. He writes:

> In the case of cogito-like judgments, the object, or subject matter, of one's thoughts is not contingently related to the thoughts one thinks about it. The thoughts are self-referential and self-verifying. An error based on a gap between one's thoughts and the subject matter is simply not possible in these cases. When I judge: I am thinking that

[8] To make things even more complicated, the [hereby] in brackets also matters. As Borgoni (2018) self-ascribing a judgment can also be non-cogito-like if the current ascription and past judgment are distinct mental occurrences. 'I judge that there will be no third world war' can be non-cogito-like (Borgoni, 2018, p. 683).

> writing requires concentration, the cognitive content that I am making a judgment about is self-referentially fixed by the judgment itself; and the judgment is self-verifying. (Burge, 1988, p. 658)

The important point here is that the content of the judgment is self-referentially fixed by the judgment. The question whether the judgment is about *p* or *p** does not arise, because no identification of the content takes place. There is no distinct judgment about the first-order judgment that might go wrong and hence no room for error. Borgoni makes this explicit with a condition for cogito-like judgments:

> [T]he second-order judgment and the ascribed mental occurrence both are part of a single thought. The self-ascribed mental occurrence is thought in and through the performance of the second-order judgment. (Borgoni, 2018, p. 683)

Whenever this is the case it does not matter whether I fully understand the content of the judgment because whatever the first-order content is, the same token content is used in the second-order judgment. Both first-order mental occurrence and second-order judgment are a single thought, therefore the content involved is the same. This also guarantees that there are no brute errors in cogito-like judgments.

But importantly, this only goes for cogito-like judgments. My judgment that 'I [hereby] judge that Los Angeles is at the same latitude as North Africa' is true even if I do not fully understand what 'North Africa' is. But non-cogito-like judgments function differently. My self-ascription 'I believe that Los Angeles is at the same latitude as North Africa' is different because it involves a judgment about a standing belief. It involves two different thoughts: a first-order belief and a distinct second-order judgment. Here the second-order judgment needs to pick out the right first-order belief. And it ought to do so safely if we accept a safety condition for knowledge. If

this is right, cogito-like judgments make for self-knowledge regardless of whether one has full understanding of the content of a belief. But this does not threaten my argument, because it still applies to non-cogito-like self-knowledge – and most self-knowledge of propositional attitudes seems to be non-cogito-like.[9]

Generalising the Results

I have shown that FULL UNDERSTANDING follows from Stalnaker's conception of propositions and a sufficiently strong safety condition for knowledge. Moreover, I argued for UNDERSTANDING THRESHOLD, the principle that at least some threshold for understanding follows for weaker accounts of safety. One might suspect that we could just reject one of the two assumed views altogether, either Stalnaker's conception of propositions or the safety condition and thereby avoid any substantial understanding requirements. But very similar arguments can be made for related notions of a proposition and conditions for knowledge. For instance, the same work that the safety condition does in the argument could also be done by a condition of being able to rule out relevant alternatives as proposed by Dretske (1970), and the same work that Stalnaker's account of propositions does can also be done with an identification of propositions with truthmaker conditions (Jago, 2017). I conjecture that any knowledge condition that rules out errors due to luck and any account of propositions that allows for two similar propositions that a subject cannot tell apart will end up in a version of FULL UNDERSTANDING or UNDERSTANDING THRESHOLD. Lack of understanding of *p* that

[9] I want to note that Burge also takes non-cogito-like self-knowledge to be authoritative and usually immune to brute errors. His argument for non-cogito-like self-ascriptions is transcendental and quite different from the one for cogito-like self-ascriptions (cf. Burge (1996)). I will bracket this argument here. Interestingly, he does suggest that the entitlement to self-knowledge does rely on understanding at least to some degree and differs from the cogito-like cases. He writes:

> The person's epistemic entitlement to the self-ascriptions presupposes understanding. Understanding is, as I have noted, dependent on and local to causal-perceptual relations to a given environment. But the entitlement that underlies knowledgeable cogito-like thoughts and other self-ascriptions does not seem local and seems to survive such switches. (Burge, 1996, p. 97)

makes it impossible for an agent to tell *p* and *p** (and *p***…) apart brings in relevant luck to the agent's belief that the agent believes that *p*. And any account of knowledge that rules out luck is therefore incompatible with this lack of understanding *p*. So simply rejecting Stalnaker's view or a safety condition are not easy ways out. I do not have an argument proving this conjecture and not enough space to go through all possible candidates for relevant accounts of knowledge. I can, however, show how my style of argument can be applied to one more popular account of knowledge: Sosa's virtue reliabilism (Sosa, 2007; 2010; 2015). In Sosa's picture a belief has to be formed aptly to amount to knowledge. That is, it needs to be accurate because of the agent's adroitness – because of the agent's competence or ability. Instances of mere lucky belief are ruled out because the truth of such a belief is not the result of the agent's competence. Self-knowledge of belief in Sosa's picture requires a competence to detect one's own beliefs, such that one's belief about one's own belief is true in virtue of that competence.

Sosa's virtue reliabilism does not entail safety as used in my initial argument, but a restricted version of safety relativized to his SSS account of abilities that is built into the aptness of a belief. Successfully exercising an ability requires that one manifests the seat of the ability when one is in a suitable situation and an appropriate shape. Greco has summarised this restricted safety in Sosa as follows:

> *SSS-relative Safety* A belief is SSS-relative safe just in case: In close worlds where S believes p from inner skill Sk, while in shape Sh and Situation Si, p is true. (Greco, 2020, p. 5152)

To illustrate the components of SSS-relative safety consider Sosa's standard example of an archer. In order to hit the target because of his ability, the archer needs to have the required

archery skill. But even if the archer has trained for years and acquired the skill, if the situation is very unfavourable, that skill cannot manifest. An archer cannot hit the target in a tornado. The winds are simply too strong to allow the archery skill to manifest. Hence, the situation needs to allow such a manifestation of the archer's skill. Moreover, even in a favourable situation, a skilled archer might not be able to manifest that skill if the archer is in a bad shape. A drunken archer will not be able to skilfully hit the target. Being drunk is a shape that works against manifesting the inner skill, even if the environment would be favourable. Only in the right shape and the right situation the archer can manifest their inner skill and hit the target because of the archer's ability.

SSS-safety applies this picture to the production of beliefs. In order to determine SSS-safety one only investigates whether the production of the belief is safe given the right shape and situation. If one's exercise of an ability in favourable situations and in a favourable shape can easily lead to false beliefs, then SSS-safety is not met and the agent lacks the competence required for apt belief. A belief that could have been easily false because of one's ability is not adroit enough and therefore not apt. Now all the building blocks are in place to provide an argument that self-knowledge of a belief that *p* in the virtue reliabilist framework also comes with requirements of understanding *p*.

1_A') A subject S has an ability that generates a belief that S believes that *p*.

2_P') An aptness requirement: a belief can only amount to knowledge if it was generated aptly, that is, if it is accurate because of S's ability.

3_P') A modal understanding of propositions: A proposition is understood as a set of possible worlds in which the proposition is true.

4_P') For any non-trivial proposition *p* there is a very similar proposition *p**, such that without full understanding of p S would not be able to tell *p* and *p** apart.

5_A') Suppose that S lacks full understanding of *p*.

6_P') If a subject S cannot tell *p* and *p** apart due to a lack of understanding, then S's exercise of the ability that formed the belief that *p* could have easily led to a belief that *p** instead.

7_C') S's exercise of S's ability generated a belief that S believes *p* that could have easily been false, because S could have been believing *p** instead of believing *p* and would not have noticed the difference. (from 4_P', 5_A', 6_P')

8_C') S's belief that S believes that *p* is not apt, because S's exercise of the ability could have easily resulted in a false belief. The belief that *p* is not true because of S's ability. (from 2_P', 7_C')

9_C') S does not know that S believes that *p*, but merely believes that S believes that *p*. (from 1_A', 2_P', 8_C')

Again, as with the original case, premise (6_P') is needed to bridge the gap from (4_P') and (5_A') to (7_C'). If S cannot tell *p* and *p** apart then the ability cannot be sensitive for any difference between *p* and *p**. With an ability that is not sensitive to such a difference little needs to change for the ability to generate *p** instead of *p*. Given the world similarity considerations from above a world in which S believes *p** is very close. This establishes (6_P') and leads to (7_C'). Because of S's lack of understanding S's ability is also unable to differentiate between believing *p* and believing *p**. Hence, the ability might easily lead to S believing that S believes *p*, even though S believes *p**, or the other way around. If this is right, then S's belief that S believes *p* will always be true by luck, if it is true. It will not be true because of the ability, given

that the ability cannot ensure the truth even in favourable situations and the subject being in a favourable shape. If the belief is not true because of the ability, it is not apt, and hence, not knowledge.

One might respond here that the ability does not need to be perfectly reliable in order to satisfy the aptness condition. Reliable enough will do. If the virtue reliabilist chooses this path, then the considerations for weaker safety accounts kick in. The virtue reliabilist will always need some amount of understanding *p*. Perhaps not full understanding, but nevertheless a very good understanding of *p*.

Conclusion

I have shown that self-knowledge of belief requires sufficient understanding of the relevant proposition. How much understanding is required depends on how luck is ruled out. Strong safety accounts lead to a requirement of full understanding. Weaker safety accounts lead to a threshold of understanding. That result likely generalises for other accounts of knowledge that rule out luck and other accounts of propositions. The same style of argument can be given for Sosa's virtue reliabilism.

With the argument in hand I can follow at least three different paths. One path brings me to suspect that people have a lot less self-knowledge than one might have initially thought. There are many propositions one only partially understands and still believes. If this is right, then my argument shows that many beliefs one has cannot be known by oneself. And it likely does not stop at beliefs, but applies to all propositional attitudes. The second path is to argue that one usually does fully or sufficiently understand the propositions that one believes. If self-knowledge is common and relatively easy, and self-knowledge requires full or sufficient understanding, then perhaps one has such understanding for most of one's belief contents. If

so, then my argument shows that belief contents are very well understood. The third path is to accept very low standards for safety – at least in most context of self-ascribing belief. If we accept very weak forms of safety, then the argument does not threaten self-knowledge. However, especially weak versions of safety are unattractive as they do not sufficiently rule out a problematic form of epistemic luck. That is, if the safety condition becomes too weak, it cannot do the job it is meant to do anymore. I am unsure which path is preferable, but either way I end up with an interesting result about the relation of self-knowledge, ruling out luck and understanding belief contents.

References

Bernecker, S. (2020). Against global method safety. *Synthese, 197*(12), pp. 5101-5116.

Blome-Tillman, M. (2020). What Shifts Epistemic Standards? DeRose on Contextualism, Safety, and Sensitivity. *International Journal for the Study of Skepticism, 10*, pp. 21-27.

Borgoni, C. (2018). Basic self-knowledge and transparency. *Synthese, 195*, pp. 679–696.

Boyle, M. (2011). Transparent Self-Knowledge. *Proceedings of the Aristotelian Society Supplementary, LXXXV*, pp. 223-241.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy, 4*(1), pp. 73-122.

Burge, T. (1988). Individualism and Self-Knowledge. *The Journal of Philosophy, 85*(1), pp. 649-663.

Burge, T. (1996). Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society, New Series, 96*, pp. 91-116.

DeRose, K. (1992). Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research, 52*(4), pp. 913-929.

DeRose, K. (2017). *The Appearance of Ignorance: Knowledge, Skepticism, and Context* (Vol. 2). Oxford: Oxford University Press.

Dretske, F. (1970). Epistemic Operators. *Journal of Philosophy, 67*(24), pp. 1007-1023.

Greco, J. (2020). Safety in Sosa. *Synthese, 197*(12), pp. 5147-5157.

Grundmann, T. (2020). Saving safety from counterexamples. *Synthese*(197), pp. 5161–5185.

Jago, M. (2017). Propositions as Truthmaker Conditions. *Argumenta, 2*(2), pp. 293-308.

Lewis, D. (1973). *Counterfactuals.* Oxford: Blackwell.

Mortini, D. (2022). A new solution to the safety dilemma. *Synthese, 200*(2).

Pritchard, D. (2005). *Epistemic Luck.* Oxford: Clarendon Press.

Pritchard, D. (2007). Epistemic Luck. *Synthese*(158), pp. 277-298.

Pritchard, D. (2012). Anti-Luck Virtue Epistemology. *Journal of Philosophy, 109*(3), pp. 247-279.

Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives, 13*, pp. 141-153.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge* (Vol. 1). Oxford: Oxford University Press.

Sosa, E. (2010). *Knowing Full Well.* Princeton, NJ: Princeton University Press.

Sosa, E. (2015). *Judgment and Agency.* Oxford: Oxford University Press.

Stalnaker, R. (1984). *Inquiry.* Cambridge, MA: MIT Press.

Stalnaker, R. (1999). Assertion. In R. Stalnaker, *Context and Content* (pp. 78-95). Oxford: OUP.

Acknowledgements

An earlier version of the paper was presented at a research seminar led by Raphael van Riel at the University of Duisburg-Essen. Thank you to all participants for their feedback. Further thanks to reviewers for their helpful suggestions.

Funding

Open Access funding enabled and organized by Projekt DEAL. Work on the article has been funded by Deutsche Forschungsgemeinschaft (Project 462399384).