Transparency and the KK Principle

Nilanjan Das (MIT) and Bernhard Salow (Trinity College Cambridge)

Penultimate Draft – Please Cite Published Version

Abstract

An important question in epistemology is whether the KK principle is true, i.e., whether an agent who knows that *p* is also thereby in a position to know that she knows that *p*. We explain how a "transparency" account of self-knowledge, which maintains that we learn about our attitudes towards a proposition by reflecting not on ourselves but rather on that very proposition, supports an affirmative answer. In particular, we show that such an account allows us to reconcile a version of the KK principle with an "externalist" or "reliabilist" conception of knowledge commonly thought to make that principle particularly problematic.

The KK principle states that someone who knows that *p* is in a position to know that she knows that *p*. In addition to an enviable pedigree of historical supporters,[1] this thesis has considerable intuitive appeal. For, to put it roughly, if the KK principle is false, rational agents can be alienated from their own attitudes and actions in a counterintuitive manner. One way to bring this out is by noting that there seems something self-undermining or incoherent about someone who says (in thought or out loud) something of the form "while it is raining, I'm not willing to take a stance on whether I know that it is." But if nothing in the vicinity of the KK principle is correct, this is hard to explain. For if there are counterexamples to KK, there are fully coherent agents who know *p* without being in a position to know that they know this. Plausibly, such agents would be justified (at least sometimes) in judging and asserting that *p* while refusing to take a stance on whether they know that *p*. In other words, they would be justified in making the self-undermining or incoherent judgements described above.[2]

[1] Hintikka (1962) takes Plato, Aristotle, Augustine, Averroës, Aquinas, Spinoza, Schopenhauer, and Prichard to have defended the similar, but stronger, principle that if an agent knows that *p*, she also knows that she knows that *p*. Williams (1978) also attributes this thesis to Descartes. However, this thesis might be too strong: if knowledge requires belief, this entails that an agent who knows that *p* always already believes that she knows that *p*, which is implausible on many conceptions of belief. That is why we prefer the weaker version of the principle which only says that knowing entails *being in a position to know* that one knows.

[2] See McHugh (2010, p. 244), Cohen and Comesaña (2013, pp. 24-25), and, especially, Greco (2015a, Section 6, 2015b) for more careful arguments along these lines. Matters are complicated here: Benton (2013) and Marušić (2013) argue that observations of this sort can be explained without relying on KK. Lacking the space to discuss

Despite its considerable appeal, the KK principle is highly controversial, and much debated in the recent literature.[3] In general, however, these debates have ignored the question of how it is that an agent who knows that *p* would come to know that she knows this.[4] This is unfortunate since, on the face of it, the second issue seems to bear on the former. If one discovers what one knows by inference from one's behaviour or via some kind of 'inner eye', it would be surprising if facts about what one knows (unlike virtually any other kind of fact one might learn in these ways) were always within reach. By contrast, if one could somehow discover that one knows that *p* by inference from one's knowledge that *p*, as suggested by certain 'transparency' accounts of self-knowledge, KK looks quite similar to an attractive closure principle stating that agents are in a position to know the obvious consequences of the claims they know.

In this paper, we spell out and defend the thought that a transparency account of self-knowledge supports the KK principle. We begin, in section 1, by summarizing what a transparency account of self-knowledge might look like. In section 2 we show how the transparency account predicts certain similarities between KK and a plausible closure principle for knowledge. In section 3, we bring out a potential disanalogy, arising from the safety of beliefs about one's knowledge, which threatens to show that KK fails even on a transparency account. In sections 4 and 5, we isolate what we take to be the key question for evaluating this threat: how should one specify the basis of an inferential belief, the method by which it was formed? We argue that one plausible answer to this question defuses the threat, and that other natural answers (which would not defuse the threat) should be rejected on independent grounds. In section 6, we respond to an objection to our argument. Finally, in section 7, we zoom out a little to explain why we should, quite generally, expect the transparency account to make KK compatible with the reliability condition on knowledge.

---

such responses, we offer these observations not as a watertight argument for KK, but only to provide a sense of the kinds of considerations making that principle attractive.

[3] Alston (1980), Feldman (1981), and Williamson (2000) offer powerful objections to KK. In response, there has also been a resurgence of KK-defenders, such as Stalnaker (2009, 2015) and Greco (2014).

[4] A notable exception is McHugh (2010), who also discusses its relevance to the KK principle.

1. Transparent Mental States

A mental state is *transparent* if an agent can come to know or justifiably believe that she is in that mental state by attending to the states of the world that the mental state in question is about. Several writers defend the transparency of beliefs and desires.[5] In relation to belief, for example, Gareth Evans famously claims that an agent can discover that she believes that there will be a third world war, just by reflecting on whether there will be a third world war.

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, p. 225)

Defenders of this account have often claimed that such world-directed reflection can generate empirical *warrant* or *justification* for beliefs about the agent's own mental states.[6] However, in order to establish that such a procedure could generate *knowledge*, we would have to show that it could yield *non-accidentally true* beliefs about the agent's own mental states. Alex Byrne (2005) explains in some detail how that could be the case.

Following Byrne, let us say that an agent *follows* a rule of the form, "If condition C holds, then φ" just in case she φ-s because she recognizes, and therefore *knows*, that C holds. Let us also say that an agent *tries to follow* a rule of this form if she φ-s because she *believes* that C holds.[7] This distinction can easily be applied to an inferential rule like OR.

> OR. If *p*, believe that *p* or *q*.

[5] See Evans (1982), Dretske (1994), Gallois (1996), Moran (2001), Byrne (2005, 2012) and Fernandez (2013).

[6] For example, both Dretske (1994) and Fernandez (2013) are concerned with the question of justification, and not of knowledge.

[7] It is crucial for this conception of rule-following to assume that, in order to follow or try to follow a rule, an agent need not independently know or even believe that she is following or trying to follow the relevant rule. Otherwise, in order to follow the rule "If condition C holds, then φ", the agent would have to antecedently know, or at least believe, that she believes or knows that C holds. But that would defeat the purpose of having a rule like BEL or KNOW, which we discuss below. However, this assumption is quite natural. After all, we often follow rules of grammar without independently knowing that we are following those rules.

An agent *tries to follow* OR if she comes to believe that *p* or *q* by inferring it from her belief that *p*, and *follows* the rule if her belief that *p* also amounts to knowledge.

Byrne then asks us to consider the following inferential rule.

BEL. If *p*, believe that you believe that *p*.

Clearly, any belief formed by trying to follow BEL (i.e. any belief formed via the inference "p; therefore I believe that p") will be true. For to try to follow BEL, an agent must believe that *p*; and if the agent believes that *p*, the belief that she believes that *p* (i.e. the belief she formed by trying to follow BEL) is true. But this means that, when an agent comes to believe that she believes that *p* by trying to follow BEL, the agent couldn't have formed a false belief in the same way. Moreover, in many cases the agent will be such that she couldn't easily have formed a belief that she believes that *p* in some other way.[8] On plausible accounts of non-accidentality, these two conditions are sufficient to make the truth of her belief that she believes that *p* non-accidental. This will typically be enough to make her belief knowledge. So, often, when an agent believes that *p*, she can come to know that she believes that *p* by trying to follow BEL. This, says Byrne, explains how an agent can often come to know that she has a belief that *p*, by using an inference whose only premise is *p* itself. It thus explains the transparency of belief.[9]

[8] Byrne (2005) allows for cases in which this second condition fails to hold, e.g. when the agent could easily have formed the (false) belief that she believes that *p* because she has had "too much coffee" (p.97). In those cases, Byrne maintains, the belief formed by following or trying to follow BEL is only accidentally true, and hence not knowledge. For reasons we discuss in section 4, we maintain that the nearby possibility of forming a false belief with the same content *on a different basis* does not render the truth of a belief non-accidental. If that's right, and (as we also think) the basis of a belief formed by trying to follow BEL is *trying to follow* BEL (or, in some cases, *following* BEL) then the truth of beliefs formed by trying to follow BEL is always, not just usually, non-accidental.

[9] Byrne's view is controversial, and we can hardly defend it here. There are, however, three objections which help to further clarify the view. The first, noted by Gertler (2011), is that Byrne's account cannot distinguish knowledge of newly formed beliefs from knowledge of previous beliefs. For example, when the question, "Do you believe that *p*?" is posed, the agent might attend to the evidence bearing on whether *p* and come to believe that *p*. Then, by following BEL, she could come to know that she believes that *p*. But it is compatible with this that the agent previously did not believe that *p*. So, the procedure described by Byrne may not yield knowledge about beliefs that obtained when the question was posed. In response to this worry, Byrne (2011, p. 208) points out that his version of the transparency proposal doesn't maintain that one can determine that one believes that *p* by reflecting on the evidence regarding the claim that *p*. It may well be that such reflection would create a belief one didn't hold previously, or undermine one that one held initially. The transparency proposal maintains only that there is at least one route available to an agent who believes that *p* which would enable her to come to know that she believes this - namely, trying to follow the relevant instance of BEL, which (perhaps unlike the process Evans describes) does not make reference to evidence about the claim that *p*.

Byrne (2012) observes that the account could also explain the transparency of knowledge. For consider the analogous inferential rule:

KNOW. If *p*, believe that you know that *p*.

Take an agent who comes to believe that she knows that *p* by *following* KNOW. Her belief that she knows will definitely be true, since she wouldn't count as following the rule unless she knew that *p*. Moreover, it seems plausible that the truth of the belief will usually be non-accidental. Often, this will be sufficient to make the belief knowledge. So, often, an agent can come to know that she knows, by following KNOW.

In what follows, we assume that following KNOW is a method we use to learn about our knowledge. How natural is this assumption? Sometimes, at least, we do answer the questions about our own knowledge by attending to facts about the external world. For example, when I go on holiday, I may wonder whether I should take my address book in case I want to write postcards. I go through my friends one by one (X lives at A, Y lives at B, etc.) and conclude that I know where everyone lives, so that the address book won't be necessary. In doing so, I settle a question about my knowledge by attending to a state of the external world. This, of course, is not to say that following KNOW is the only way of settling such questions. It is enough for our

---

The second worry, raised by Shoemaker (2009), is this. A rule enjoins us to *perform an act*. So, it is not clear whether we can ever comply with a rule like BEL; after all, a belief is not an act, but a standing state. To us, this worry seems to target an inessential detail of how Byrne (2005, 2012) formulates the view. Instead of stating BEL as an imperative for agents to follow, we could reformulate it as an inference scheme, as Byrne (2011), following Gallois (1996), does: "*p*. Therefore I believe that *p*." Clearly agents have some way of forming beliefs in line with inference schemes, even if doing this does not qualify as an action.

The third objection, also raised by Shoemaker (2009) and elaborated by Boyle (2011), says that rules like BEL are not inferential rules that a reasonable person can follow, because the premise in BEL cannot reasonably be taken to be evidence for the conclusion. We agree that BEL (and KNOW, which is the more relevant case for us) has the feature Boyle identifies; but we are not sure that this makes following it unreasonable. After all, these rules have other good-making features, since the beliefs they generate are non-accidentally true. Moreover, there are many situations, such as the address book example described below, in which reasoning in accordance with these rules strikes us as very natural.

We should also note that, while we will formulate our discussion in terms of Byrne's version of the transparency account, some of the core insights might be available to transparency theorists more generally; see section 7 for discussion.

purposes for it to be the case that following KNOW is *one* method, amongst others, of gaining knowledge about our own knowledge.

2. Hope

Byrne is content to claim that the possibility of following a rule like KNOW *sometimes* puts us in a position to know that we know. One might, however, hope for more.

Following a rule like KNOW is importantly similar to following a deductive inference rule like OR, which says, "If *p*, believe that *p* or *q*." To see the similarity, compare OR to the following rule discussed by Byrne (2005):

DOORBELL. If the doorbell rings, believe that someone is at the door.

OR and DOORBELL are quite different. An agent follows OR just in case she believes that *p* or *q*, because she knows that *p*. So, a belief formed by following OR is always true. By contrast, following DOORBELL need not always yield true beliefs. Sometimes, an agent can know that the doorbell has rung, even though there's no one at the door; a wiring defect might have made the doorbell ring. It should be obvious that KNOW is, in this respect, much more similar to OR than to DOORBELL: a belief formed by following KNOW is always true. To put the point slightly differently, the inferential transitions prescribed by OR and KNOW have a common virtue which that prescribed by DOORBELL lacks: they will never take you from a completely flawless belief (that is: a piece of knowledge) to a false one.

It is natural to think that the possibility of reaching a conclusion by applying simple deductive rules like OR from premises we know is *always* sufficient to put us into a position to know that conclusion. Given the similarity between following KNOW and following deductive rules like OR, it is thus tempting to think that transparency similarly puts us into a position to

know that we know not just *sometimes* but *always*. In other words, one might hope that the transparency of knowledge will support KK:[10]

> KK. If an agent knows that *p*, then she is in a position to know that she knows that *p*.

Before it can become defensible, however, this hope must be somewhat qualified. After all, consider the equally unrestricted principle *Closure*:

> *Closure*. If an agent knows that *p*, and *q* is an obvious logical consequence of *p*, then the agent is in a position to know that *q*.

In their unrestricted forms, both principles seem susceptible to at least three kinds of counterexamples.

First, the agent in question might not have the concepts required to believe the claim in question. In the case of *Closure*, the agent might lack one of the concepts involved in q but not p; in the case of KK, the agent might (like Castaneda's (1979) Externus) have no self-concept, or might lack the concept 'knowledge'. In such a scenario, the unrestricted principles seem to fail.

Second, the knowledge that *p* might not be *inferentially accessible*, i.e., accessible for making the relevant kinds of inferences.[11] For example, someone might know that a friend's

[10] Dokic and Égré (2009) and McHugh (2010) also defend KK on the basis of transparency. We will defend two specific theses: that the transparency account motivates a tight analogy between *KK* and *Closure* and that beliefs formed by following KNOW always meet the safety requirement on knowledge. Neither of these is anticipated by these two papers: the analogy with *Closure* is absent from both, and while the two papers discuss the issue of safety, their treatment differs significantly from ours. McHugh (2010, pp.251-252) grants that an agent may know without being able to form a safe belief that she knows, and grants that safety may be a necessary condition on knowledge. However, he argues that such an agent would nonetheless in an interesting sense be 'in a position to know' that she knows, since this needn't be understood as being able to form a belief that would amount to knowledge. He may be right that there is such a sense; but if our argument works, it shows that agents who know are in a position to know that they know also in a more robust sense which does require being able to form a belief that would amount to knowledge. It is less clear to us what exact view Dokic and Égre (2009) take on the safety requirement. They maintain that higher-order reflective knowledge does not require us to leave a margin for error; but it is unclear whether they think that this is because such reflective beliefs needn't be safe to be knowledge or because such beliefs can be safe even when they leave no margin. Our argument, if successful, offers a way of substantiating this second version of their view.

phone number is 617-785-6252, while being unable to access that information in any way other than by dialling the number without thinking about it (and being unaware that he has this ability). *Closure* would then predict that the agent is in a position to know that the friend's number contains two 6s, while KK predicts that the agent is in a position to know that he knows his friend's number. In both cases, however, this conclusion looks somewhat counterintuitive.

Third, there are cases in which contemplating the proposition the agent is supposedly in a position to know would generate worries that, in some way or another, would undermine the agent's belief in the premise and thus prevent her from drawing the inference. For example, while I know that I will be teaching logic next year, I might not be able to deduce from this that I won't die in a traffic accident beforehand. For, if I were to consider the possibility of dying in a traffic accident, the uncertainty about this possibility would make me reconsider my conviction that I will teach logic next year.[12] Thus, even though I know that I will teach logic next year, I cannot use that knowledge to learn that I won't die in a traffic accident.

Similar counterexamples also seem to occur in the case of KK. I remember that Germany won the 2014 world cup. But, when I think about whether *I know* that Germany won the 2014 World Cup, I might feel the need to dig further: what exactly do I remember, how reliable are those memories, might I have been misled?[13] (Though it's important for our purposes that, as the earlier address example brings out, we do not always feel such a need.) If I do this, there is no guarantee that I will conclude that I do know, even if, initially, I did. But, much like in the

---

[11] Saying that a belief is *inferentially accessible* (usable in inference) is different from saying that it is *reflectively accessible* (i.e. that the agent is in a position to know that she has this belief). Given the transparency account, it might turn out that inferential accessibility entails reflective accessibility (since we can use the transparency inference to learn of an inferentially accessible belief that we have that belief); but this connection is hardly an obvious or uncontroversial one.

[12] Cf Nagel (2011) who suggests that, when an agent contemplates the proposition about the traffic accident, she switches to a reflective mode of cognition and is unable to endorse her prior judgement about teaching logic next year, or base her judgements about the traffic accident proposition on that prior judgement. One might want further explanation for why agents often become unable to endorse the relevant belief in these circumstances. *Contextualists* (such as Cohen (1988, 1998), Lewis (1996), Neta (2002), and Rieber (1998)) might appeal to the fact that contemplating the propositions in questions raises new error possibilities, which change what the agents in question mean by 'knowledge', and that it is clear to subjects that their belief does not meet those new standards. *Subject-sensitive invariantists* (such as Hawthorne (2004, ch.4)) might appeal to the fact that considering these propositions makes salient new decision situations, ones in which the stakes would be high enough to prevent the relevant belief from counting as knowledge. For our purposes, it does not much matter what exactly the explanation is, provided that (as seems plausible) it also applies in the counterexamples to KK.

[13] Thanks to Jennifer Nagel for this observation.

analogous counterexamples to *Closure* where it seems weird to say "Sure, I will teach logic next year, but will I die in a traffic accident before then?", saying "Sure, Germany won, but do I know this?" seems a weird reaction here. This suggests that, when this kind of reflective process is triggered, I must drop the belief in the premise (that Germany won) as well, or at least mustn't "endorse" it. In this sense, then, the counterexamples are analogous, both arising from the fragility of knowledge under reflection.

Importantly, these problems do not, we think, tell against the thought that some qualified but still interesting version of *Closure* is correct. Perhaps, we can restrict *Closure* in the following manner to avoid these problems.

> *Restricted Closure.* For any agent who is able to apply the relevant deductive rules to the premise that *p*, if she knows that *p*, and *q* is an obvious logical consequence of *p*, then she is in a position to know that *q*.

The phrase "is able to apply the relevant deductive rules to the premise that *p*" functions as a placeholder here, and we will not spell it out further. However, the previous discussion shows that the following three are necessary conditions for the agent to be "able to apply the relevant deductive rules to the premise that *p*": (1) the agent must possess all the relevant concepts; (2) the agent's knowledge of the premise must be inferentially accessible; (3) the agent must be able to retain her knowledge that *p* when the issue of whether *q* is true becomes salient.

The realistic hope is then that the transparency account of how we know that we know could show that optimism about a suitably qualified version of KK, like the following, is no more futile than optimism about *Restricted Closure*.

> *Restricted* KK. For any agent who is able to apply KNOW to the premise that *p*, if the agent knows that *p*, she is in a position to know that she knows that *p*.

Once again, the phrase "is able to apply the relevant deductive rules to the premise that p" functions as a placeholder, for which we offer no sufficient conditions. This does not, however, render the principle trivial. For the contention is that *Restricted* KK will be true when this placeholder is understood in whatever way it needs to be understood to render *Restricted Closure*

true. Given what we said above, however, this claim is weak enough to not be refuted by counterexamples to KK, such as the ones discussed above, that rely on agents who fail to possess the relevant concepts, whose knowledge that *p* is inferentially inaccessible,[14] or who would lose their knowledge or belief that *p* when the question of whether they know that *p* becomes salient.

Before moving on, it's worth quickly pointing out that *Restricted* KK is strong enough to explain the data motivating KK that we appealed to in the introduction. One way to see this is by noting that agents who violate KK for the reasons discussed above couldn't make judgments of the kind "*p*, but I take no stance on whether I know that *p*". The analogy between KK and *Closure*, however, gives us a more general way of verifying this. For *Closure* is also naturally motivated by the badness of certain judgments or assertions: when *p* is an obvious consequence of *q*, there is something very odd about someone who judges or asserts that *p* while refusing to take a stance on whether *q*.[15] This means that, however the restriction in *Restricted Closure* is understood, the resulting principle had better still explain why rational agents do not make such judgments or assertions. But then *Restricted KK* will explain why rational agents don't judge or assert "*p*, but I take no stance on whether I know that *p*" in an exactly parallel fashion.

3. Despair

Unfortunately, the analogy between *Closure* and KK doesn't take us all the way. For there is also an important difference between KNOW and rules like OR. Applying OR takes an agent from the belief that *p* to the belief that *p* or *q*; if the first of those beliefs is true, so is the belief formed in the inference. By contrast, applying KNOW takes her from the belief that *p* to the belief that she knows that *p*; since not every true belief is knowledge, it is possible that, even though the belief in the premise is true, the belief formed via the inference is not. The inferential transition prescribed by OR therefore has a further virtue not shared by the one prescribed by KNOW: it will never take you from a true belief to a false one.

[14] Note that the distinction between reflective accessibility and inferential accessibility---discussed in footnote 11---plays an important role here. Requiring the knowledge to be reflectively accessible (i.e. requiring that the agent be in a position to know that she has this knowledge) would trivialize the principle. However, our principle requires only that the knowledge be *inferentially* accessible; the principle thus makes a substantive claim.

[15] For versions of this observation, see DeRose (1995) and Hawthorne (2004).

This seems to matter, if we think that safety is a necessary condition on knowledge: you know that *p* only if you couldn't easily have falsely believed that *p* (i.e. don't falsely believe that *p* in any nearby possible worlds).[16] For suppose you know that *p*. Then, by safety, you don't falsely believe that *p* in any nearby worlds. So applying OR to form the belief that *p* or *q* will yield true beliefs not just in this world, but also in any nearby worlds. So the belief that *p* or *q* will be safe as well as true. There is no analogous guarantee that a belief that one knows, formed by applying KNOW, will be safe. For knowing that *p* guarantees only that *p* is true in nearby worlds in which it is believed, not that it is known there. So there may be nearby worlds in which you apply KNOW to go from the true belief that *p* to the false belief that you know *p*. Your actual belief that you know, while true, would thus seem to be unsafe.

Timothy Williamson's influential criticism of the KK principle nicely illustrates this worry.[17] Consider Mr Magoo, who is taking part in a contest of judging the height of randomly selected trees. Mr Magoo's ability to judge such heights is good but imperfect. In particular, if Mr Magoo actually judges that tree T is at least *x* inches tall, he could easily have made that same judgment about any tree up to 5 inches shorter. This means that, when faced with a 100 inch tree, the strongest claim Mr Magoo can know is that he is faced with a tree that is at least 95 inches tall. For suppose he were to believe that he is faced with a tree that is at least, say, 98 inches tall. Then his belief would be unsafe. For he could easily have believed the same thing when faced with a tree 5 inches shorter; and, since such a tree would have been only 95 inches tall, his belief would then have been false.

None of this raises trouble for *Closure*: since the belief that the tree is at least 95 inches tall is safe, anything deduced from it is true in all nearby worlds in which it is inferred from this belief, and hence equally safe. But the example does raise trouble for KK. For suppose that,

[16] This gloss on safety is inadequate: when the claim that *p* couldn't easily have been false (e.g. because it is a necessary truth), this definition predicts that an agent's belief that *p* is safe no matter how unreliably she forms that belief. A better gloss is that S's belief that *p* is safe only if S couldn't easily have formed a *relevantly similar* false belief (which needn't be a belief that *p*). But the issues we're interested in here doesn't depend on considerations about propositions that couldn't easily have been false. So, we will set this complication aside.

[17] See Williamson (2000, chapter 5). We will present a simplified version of the argument, since the additional details aren't relevant to our particular concerns. Our response would block the original argument in exactly the way indicated by Dokic and Égré (2009).

faced with a 100 inch tree, Mr Magoo believes that he knows the tree to be at least 95 inches tall. Then it would seem that he could easily have formed that same belief (the belief that he knows the tree to be at least 95 inches tall) when faced with a tree 98 inches tall. After all, we're assuming that he could easily have believed, in such a scenario, that the tree is at least 95 inches tall, and so he could have come to believe that he knows this by applying a rule like KNOW. But, had Mr Magoo been faced with a tree 98 inches tall, he would not have known that it is at least 95 inches tall. By the same reasoning as in the previous paragraph, the strongest thing Mr Magoo could have known about a 98 inch tree is that it is at least 93 inches tall. Thus, Mr Magoo could easily have believed himself to know that the tree is at least 95 inches tall when he knew no such thing (though not: when no such thing was true), and so his belief that he knows is unsafe. Therefore, Mr Magoo knows the tree to be at least 95 inches tall, but seems to be in no position to know that he knows this. Hence, KK is false.

4. The Basis of Inferential Knowledge

In section 2 we saw some initial ground for optimism that the transparency account of introspection would vindicate KK. But in section 3 we encountered a major challenge: even if beliefs formed by following KNOW are always true, there is no obvious reason to think that they are always safe. And if they aren't always safe, they aren't always knowledge. So it is not clear that the transparency account of introspection supports KK.

Our discussion of safety, however, was extremely informal. We used something like the following gloss: an agent's belief that *p* is safe if that agent couldn't easily have falsely believed that *p* (i.e. doesn't falsely believe that *p* in any nearby worlds). But a little reflection makes obvious that this is too stringent a requirement. Imagine that someone, following DOORBELL, comes to believe that there is someone at the door, in a situation in which the doorbell ringing is in fact a reliable indicator of someone being at the door. It seems clear that such a belief would count as knowledge. And this is so even if there are nearby worlds where the agent comes to falsely believe that there's someone at the door, not by following DOORBELL, but because someone mistakenly tells her that there's someone at the door. The natural conclusion is that the possibility of a false belief formed *on a very different basis* shouldn't render our agent's actual

belief unsafe.[18] A better gloss on safety is thus: an agent's belief that *p*, formed on basis *B*, is safe if that agent couldn't easily have falsely believed that *p* on basis *B*.[19]

Whether the transparency account supports KK against the challenge from safety depends entirely on what we take to be the basis of a belief formed by following KNOW. For consider the two obvious options. (We will consider a third in section 4, and a fourth in footnote 27.) According to the first, the basis of the belief is *following* KNOW, so that (by the definition of 'following' in section 1) no other belief counts as having the same basis unless it is also formed by applying KNOW to known premises. According to the second, the basis of the belief is *trying to follow* KNOW, so that every belief formed by applying KNOW to premises which are themselves believed (even if they aren't known) counts as having the same basis. (It's worth stating explicitly that it's not obvious that the basis of a belief formed by following a rule R is *following* R: even if a belief was in fact formed by applying R to known premises, that doesn't mean that all and only this information about it specifies its basis.)

Now, we have already seen that KNOW is a rule which, whenever it is followed, will result in a true belief. If the basis of a belief formed by following KNOW is *following* KNOW, this guarantees that following KNOW will always yield safe beliefs. For a belief in some nearby possibility will count as being formed on the same basis only if it was also formed by following KNOW; and we've already shown that such a belief must always be true. Hence the agent couldn't have formed a false belief on the same basis. So the belief is safe.

No such argument is possible if the basis of a belief formed by following KNOW is *trying to follow* KNOW. For, unlike trying to follow BEL, trying to follow KNOW is not guaranteed to yield a true belief. This is essentially what happened in the case of Mr Magoo. In the actual scenario, in which Mr Magoo is faced with a 100 inch tree, Mr Magoo follows KNOW in moving from his knowledge that the tree is at least 95 inches tall to his belief that he knows this. In the

[18] DeRose (1995) offers a superficially different response to this kind of example, simply requiring that possibilities resemble the actual world in a particular way to count as ones that 'could easily have happened' in the relevant sense. We're inclined to think that talk of bases is mostly a way of putting a label on the resemblance in question.

[19] As mentioned in footnote 16, one might want to require also that the agent couldn't easily have formed a relevantly similar false belief on basis B; we will continue to ignore that complication.

nearby scenario in which Mr Magoo is faced with a 98 inch tree and still believes that it is at least 95 inches tall, he merely tries to follow KNOW in coming to believe that he knows the tree to be at least 95 inches tall. If *trying to follow* KNOW is the basis of his actual belief, that nearby false belief counts as being formed on the same basis, and hence prevents the actual belief from being safe. By contrast, if *following* KNOW is the basis, the possibility of this false belief formed by trying to follow KNOW is irrelevant.

Which of these – *following* KNOW or *trying to follow* KNOW – is the right way of individuating the basis of a belief in fact formed by following KNOW? We doubt that we have a sufficiently firm pre-theoretic grip on the relevant notion of 'basis' to answer this question directly.[20] But there are theoretical grounds for imposing some high-level constraints, which (together with substantive judgments about which beliefs amount to knowledge) will be enough for our purposes.

One natural thought is that the manner in which we individuate the bases of beliefs within the safety-theoretic framework shouldn't be piecemeal. In other words, the basis of pre-theoretically similar beliefs should be individuated similarly. In particular, the basis of a belief formed by following an inferential rule should be individuated in the same manner, irrespective of what the rule is. More precisely,

> *Generality Constraint.* If the basis of a belief formed by following rule *R* is bearing relation *X* to *R* (e.g. following, trying to follow, etc.), then the basis of a belief formed by following rule *R′* is bearing relation *X* to *R′*.[21]

[20] Goldman (2009) points out that the notion of 'basis' ordinarily applies to mental states like experiences and memories, which do not have the global reliability properties which bases should have by lights of the safety theorist. Williamson (2009b) responds that one should accept a 'liberal conception' of bases, under which bases may include processes of belief-formation as well as facts about the causal background against which those processes operate. We agree; but a consequence of this is that we cannot rely on direct intuitions about the bases of beliefs.

[21] One might object to the *Generality Constraint*, say, because there are important disanalogies between transparent and deductive inferences (e.g. with regards to whether the premises evidentially support the conclusion), or because one holds that the notion of a 'basis' is not a theoretically tractable concept. We respond to such worries in section 6; roughly, our claim is that our response to the safety-based objection to KK retains most of its dialectical significance even if the *Generality Constraint* is rejected for such reasons.

For example, if the basis of a belief formed by following BEL is *trying to follow* BEL, the basis of a belief formed by following DOORBELL had better be *trying to follow* DOORBELL. A second constraint, linking the basis of a belief which amounts to knowledge to the explanation for why that belief is true, will be introduced and motivated in section 4. These constraints, we claim, will be enough to reduce the option space enough to make progress on the status of KK.

In particular, the *Generality Constraint* implies that if the basis of a belief formed by following KNOW is *trying to follow* KNOW, then the basis of a belief formed by following OR is *trying to follow* OR. But, as we are about to show, this predicts highly implausible counterexamples to *Closure*.

To see why, consider someone who in fact knows that there is a sheep in the field, having seen it. The sheep very nearly escaped a few seconds before our agent looked at it. Had it done so, someone would have prevented our agent from looking at the field and simply told her that there was a sheep in it. With no reason to distrust her informant, our agent would have formed the false belief. This doesn't prevent her from knowing, given how things actually proceeded, because the false belief would have been formed on a very different basis.

So far, not much of interest has happened. But now suppose that our agent applies OR to infer that there is a sheep or a cow in the field from her knowledge that there is a sheep in the field. Since OR is a paradigmatic inference rule, and our agent knows the condition, that belief should amount to knowledge. However, if we identify the basis of the belief in the disjunction as *trying to follow* OR, the belief will be unsafe. For there is then a nearby situation in which she believes the disjunction on the same basis, by inferring it from her mistaken (testimony-based) belief that there is a sheep in the field. We conclude that one shouldn't identify the basis of the belief in the disjunction as *trying to follow* OR.

The problem is naturally avoided if we take the basis of the belief in the disjunction to be *following* OR. In the nearby world in which our agent's belief in the disjunction is mistaken, that belief was not formed by applying OR to known premises. If the basis of the belief in the actual world is *following* OR, that means that the mistaken belief was not formed on the same basis. The

belief in the disjunction thus still counts as safe. This is a powerful reason to prefer thinking of the basis of the belief as *following* OR rather than as *trying to follow* OR. By the *Generality Constraint*, then, it is a powerful reason to prefer thinking of the basis of the belief that one knows as *following* KNOW rather than as *trying to follow* KNOW. And if that is how we think of the basis of beliefs formed by following KNOW, such beliefs are guaranteed to be safe.

5. Expanding Bases

However, our objection to specifying the basis as *trying to follow R* makes salient a third alternative specification of the basis. Consider again the case where our agent forms the belief that there's a sheep or a cow in the field by following OR. Here, the belief she uses when she follows OR is a perceptual belief. So, perhaps, the right specification of the inferential belief is *trying to follow* OR *using a perceptual belief*. Now, in the nearby worlds where the agent tries to follow OR using a testimony-based belief, she doesn't try to follow OR using a perceptual belief. So, the false testimony that the agent might have received in nearby worlds doesn't undermine the safety of the belief that she forms by reasoning from her perceptual belief. So the alleged counterexample to *Closure* is blocked by this third proposal.

While avoiding the counterexample to *Closure*, the third alternative still allows for counterexamples to KK. For consider, again, the case of Mr Magoo, who knows by perception that the tree in front of him is at least 95 inches tall. Now suppose that Mr Magoo follows KNOW, and comes to believe that he knows that the tree is at least 95 inches tall. If we take the right specification of the basis to be *trying to follow* KNOW *using a perceptual belief*, this belief will be unsafe. For, in the nearby case in which he takes himself to know the same thing even though the tree is only 98 inches tall, this belief that he knows will still be formed by trying to follow KNOW using a perceptual belief that the tree is at least 95 inches tall, albeit a perceptual belief that doesn't amount to knowledge. So Mr Magoo's actual belief that he knows isn't safe from error, and therefore isn't knowledge.

More abstractly, the strategy behind the 'third alternative' is this. Consider an agent who forms a belief *b* by following an inferential rule *R* using a belief that *p* which is itself held on

basis B. Then the correct specification of the basis of *b*, according to the strategy in question, is: *trying to follow R using a belief that p held on basis B*. In our examples, this makes the basis of the inferential belief something like *trying to follow OR using a belief that there is a sheep in the field held on the basis of perception* and *trying to follow* KNOW *using a belief that the tree is at least 95 inches tall held on the basis of perception*.

We have already seen that such a strategy predicts Mr Magoo to be a counterexample to KK. And we also saw that it avoids our potential counterexample to *Closure*. The second point can be generalized, to show that this strategy is compatible with *Closure* across the board. For suppose that our agent knows that *p*, and forms the belief that *p* or *q* by following OR using her belief that *p*. Let B be the basis of her belief that *p*; then the basis of her belief that *p* or *q* is *trying to follow OR using a belief that p held on basis B*. Now, any possibility in which our agent comes to believe that *p* or *q* by trying to follow OR using a belief that *p* held on basis B is, trivially, a possibility in which she believes that *p* on basis B. Since our agent's belief that *p* amounts to knowledge, any such possibility which is also nearby is one in which *p* is true. But then this possibility is also one in which *p* or *q* is true. It follows that any nearby possibility in which our agent believes that *p* or *q* on the same basis is one in which it is true that *p* or *q*; and hence it follows that her belief that *p* or *q* is safe. On this way of construing bases, then, beliefs formed by deduction from safe beliefs are themselves safe, while beliefs formed by reasoning in line with KNOW from safe beliefs needn't be. The proposal thus predicts the exact disanalogy between KK and *Closure* that has worried us since section 3.

To argue against this third alternative, we will need to delve a little deeper into the theory of bases.

According to a natural conception of *coincidence*, an event can be treated as a coincidence or an accident only if it is *inexplicable* in a certain sense.[22] To borrow an example from David Owens (1992), suppose, on a rainy day, I pray that it doesn't rain tomorrow, because tomorrow is my wedding day. Indeed, my prayer comes true. The sceptics will say that this is a mere coincidence: the fact that my prayer came true has two constituents which are independent

[22] For a similar idea, see Sorabji (1980) and Owens (1992). Sorabji traces the idea back to Aristotle.

of each other, namely the fact that I made a prayer with content C on a particular day, and the fact that C came true the next day. There is a separate explanation of why each of these happen, but there is no explanation of why they happen together. The faithful will insist that this is no coincidence: God heard my prayer and prevented the rain from continuing, so that there is an explanation of why both things happen together. According to this conception of *coincidence* or *accident*, an event is non-accidental only when there is an explanation of why all the constituents of the event happen together.

One lesson of the Gettier problem is that, when someone knows, it is non-accidental that she believes the truth. On the explanatory conception, this means that an agent knows only when there is an explanation of why the agent's belief and the truth coincide. Now, the most common motivation for the safety condition on knowledge is that the safety condition guarantees that the truth of a belief that amounts to knowledge is not an accident or a coincidence.[23] However, on the explanation-based conception of coincidence, it can do so only if we assume:

> *Explanatory Constraint.* If *B* is the basis of a belief that amounts to knowledge, then the proposition that the belief was formed on basis *B* should provide the ingredients needed to explain, together with the facts about the circumstances, why the belief is true.

For if the basis of a belief doesn't provide the ingredients to explain why the belief is true in the relevant circumstances, it is hard to see why a belief couldn't be safe even though there is no explanation at all for why the belief is true; a belief could thus be both safe and true only by accident. To prevent this result, we should endorse the *Explanatory Constraint*.

We will argue that this constraint is compatible with taking the basis of a belief formed by following OR to be *following* OR, but not something like *trying to follow* OR *using a perceptual belief*.

[23] Sosa (1999) motivates safety as a condition as an alternative to Nozick's (1981) sensitivity condition on knowledge, while Pritchard (2005) takes it to be an anti-luck condition on knowledge. In each case, the main purpose of the safety condition is to rule out instances of epistemic luck typical of Gettier-type examples.

The positive part of this strikes us as straightforward: it seems a very good explanation of why S's belief that *p* or *q* is true that she deduced it from something which she knew. The negative part is slightly trickier. Why is it not an equally good explanation that she deduced it from her belief that *p*, which she had in turn formed by perception? After all, together with the background information which shows perception to be reliable, this also entails the observation that needed to be explained.

Our worry is that the explanans is not adequately *proportioned* to the explanandum in this case.[24] Suppose I were to try explaining why a ball released on the lip of a basin ends up at the basin's lowest point, by appeal to the exact initial position and velocity of the ball.[25] Then you would have good reason to reject my proposed explanation out of hand, not because it fails to entail the explanandum, but because it brings in too many extraneous details. After all, the ball would have ended up where it did even if it had been released at quite a different part of the basin with quite a different initial velocity. Similarly, it seems to us that bringing in the basis of S's belief that *p* introduces information irrelevant to explaining why S's belief that *p* or *q* was true. After all, she would have formed the same true belief even if she had known that *p* in some other way, say by testimony.

By itself, this does not sink the proposal for individuating bases. For the *Explanatory Constraint* does not require that *all* the information contained in the fact that the belief was formed on basis *B* should feature in the explanation for why the belief is true.[26] So the proposal could be saved if there were some consequence of the fact that the belief was formed by *trying to*

[24] Some writers, like Yablo (1992) and List and Menzies (2009), take proportionality to be constraint on what counts as a cause. Others, like Weslake (2013), think that its place is instead in the theory of explanation. For our purposes, this second understanding of proportionality is enough.

[25] For this example, see Strevens (2008, pp. 434-435), who uses it to illustrate a different virtue of explanations, namely *robustness*.

[26] Moreover, we ourselves are committed to rejecting such a stronger requirement. For it seems clear that the explanation for why a belief formed by following BEL is true is not that it was formed by *following* BEL, but rather that it was formed by *trying to follow* BEL (since it would have been true, and for the same reason, either way). But the *Generality Constraint* commits us to saying that the basis of a belief formed by following BEL is *following BEL*. So not all of the information provided by the basis of this belief is relevant to explaining its truth. Nonetheless, since that the belief was formed by following BEL entails that it was formed by trying to follow BEL, and the latter is the crucial claim needed to explain why the belief is true, this treatment of BEL still satisfies the less demanding *Explanatory Constraint*.

follow OR *using a perceptual belief* which did explain why the belief was true. But what would that consequence be? Not that the belief was formed by trying to follow OR using a belief that amounted to knowledge, since that doesn't follow from the fact that it was formed by trying to follow OR using a perceptual belief.

Our best shot is that the consequence is some long disjunction, which enumerates all the ways in which the agent could have come to know that *p* in the relevant circumstances. That is, the fact we might appeal to in the explanation is that the belief that *p* or *q* was formed by trying to follow OR using a belief that was either based on perception or testimony or memory or… This disjunctive fact does seem adequately proportioned, making no mention of excess information that is irrelevant to why the belief that *p* or *q* is true.

Nonetheless, we do not think that it can feature in an explanation of why that belief is true. The problem is that there is no single uniform route from each of the disjuncts to the explanandum. For each disjunct, we need to start off explaining why the belief that p, formed in this way, would be true; and the explanation of this will be different in each case. So the disjunctive fact also cannot feature in an adequate explanation of why the belief that *p* or *q* is true. Since we see no other promising consequence which a defender of these extended bases could appeal to, we conclude that this third proposal for how to individuate bases is incompatible with the *Explanatory Constraint*.

We have now discussed three natural proposals for how to specify the basis of a belief formed by following a rule *R*:

1. The basis of a belief formed by following R is to be individuated in terms of *following* R;
2. The basis of a belief formed by following R is to be individuated in terms of *trying to follow* R;
3. The basis of a belief formed by following R is to be individuated in terms of *trying to follow* R *using a premise believed on the same basis as in the actual case.*

Furthermore, we have shown that proposals 2 and 3 are independently problematic: proposal 2 generates implausible counterexamples to *Closure*, while proposal 3 is incompatible with the

plausible *Explanatory Constraint*. Since we can see no promising fourth proposal,[27] we conclude that proposal 1 is correct. And, as we saw earlier, it follows from this that beliefs formed by following KNOW always count as safe.

6. The Role of the Generality Constraint

We have argued that, contrary to initial impressions, beliefs formed by following KNOW are always safe, since the agent couldn't easily have formed a false belief on the same basis. It may be helpful to quickly recast that argument, in a slightly different order to the way it was presented. When someone comes to believe that *q* by deducing it from such-and-such known premises, the basis of her belief is *deduction from such-and-such known premises*. This is because alternative proposals about the basis of her belief (such as *deduction from such-and-such believed premises* or *deduction from such-and-such premises believed on such-and-such a basis)* either (i) fail to guarantee that her belief that *q* is safe, thus predicting violations of *Closure*, or (ii) contradict the attractive thought that facts about the basis of a safe belief explain why that belief is true, thus violating the *Explanatory Constraint*, and threatening the connection between beliefs that are safe and beliefs whose truth is no accident. But, according to the transparency account of self-knowledge, a belief that one knows can be very much like other inferential beliefs, including deductive ones. This suggests that the bases of beliefs formed via a transparent inference should be characterized in the same manner as the bases of beliefs formed by deductive inference (we call this the *Generality Constraint*). Given what we showed about beliefs formed by deduction, this means that when an agent comes to believe that she knows that *p* by applying KNOW to a known premise, the basis of that belief is *applying* KNOW *to such-and-such known premises*. Obviously, a belief that one knows held on that basis cannot be false. Beliefs formed by following KNOW are therefore trivially safe, and the safety-based objection to KK is disarmed.

[27] Here is one more option. Let us say that an agent correctly applies the rule "if C obtains, φ" if she φ-s because she truly believes that condition C obtains. Why not take the basis of a belief formed by following *R* to be *correctly applying R*? One problem is that this proposal makes it too easy to form safe beliefs. Suppose I have a belief that *p* which is true but unsafe. Then I can use OR to form a safe belief that p or q: after all, the new belief will have been formed by correctly applying OR, so that only other beliefs formed by correctly applying OR count as being held on the same basis; and, trivially, all such beliefs are true.

Of course, opponents of KK can always respond by rejecting our premises. Most saliently, they might object to the *Generality Constraint*, perhaps because they see important disanalogies between transparent and deductive inferences (e.g. with regards to whether the premises evidentially support the conclusion), or because they object to the thought that the notion of a 'basis' is intuitively or theoretically tractable even to the extent required to motivate this very general constraint.[28] These objections raise large issues, which we cannot adequately address here. So let us instead examine what happens to the dialectic surrounding safety and KK if the *Generality Constraint* is rejected on such grounds.

We grant that, absent the *Generality Constraint*, we have provided no positive reason to accept that beliefs formed by following KNOW will always be safe. However, we have nonetheless established that the safety-based objection to KK has a gaping hole: it goes through only if we reject one very natural way of construing the basis of a belief formed by following KNOW. Moreover, it is hard to see how this hole could be plugged. If we think that transparent reasoning is so unusual that we can learn nothing about the bases of transparent beliefs by considering ordinary reasoning, or if we think that the notion of a 'basis' is a mere placeholder with little intuitive or theoretical content, how could we possibly go about determining the bases of such beliefs? Whether to accept KK will thus have to be settled on safety-independent grounds; and our suspicion (which we have, admittedly, not defended in detail) is that defenders of KK will have the edge in that debate.

One might worry that this description of the dialectical consequences of rejecting the *Generality Constraint* ignores the intuitive force of the safety-based counterexamples to KK. Many find it intuitive to say that Mr Magoo's belief that he knows that the tree is at least 95 inches tall is barred from being knowledge because that very same reasoning could so easily have led him into error. Couldn't one appeal to such a direct intuition whilst rejecting general theoretical principles like the *Generality Constraint*?

We do not think that one should. The reason is that any intuition that 'the very same reasoning' could have led Mr Magoo into error seems, to us, to rely on the more general thought

[28] Williamson (2009a, p.9-10) holds that we have virtually no grasp on the notion 'nearness' used in the safety condition, other than through our judgements about knowledge; it's natural to extend this to the notion of 'basis'.

that whether two inferences are 'the same reasoning' depends only on which beliefs they involve, or at least depends only on 'internal' facts about the agent. But our reflections on deduction (which in no way relied on the *Generality Constraint*) suggest that these more general thoughts must be rejected: whether two deductive conclusions are reached by 'the same reasoning' (in the sense relevant to safety) can depend on whether the agent knew the premises. Once one has absorbed this result, then, we see little reason to trust any intuitive appearance that Mr Magoo engages in 'the same reasoning' (in the sense relevant to safety) when he infers that he knows the tree to be at least 95 inches tall from his knowledge that it is and when he draws that same conclusion from a mere belief that the tree is at least 95 inches tall.

We thus conclude that our defence of KK against safety-based objections is fairly independent of one's view of the *Generality Constraint*. If the constraint is accepted, our arguments offer powerful reason to believe that beliefs formed by following KNOW will always be safe. If the constraint is rejected, the arguments still give us good reason to reject general arguments or intuitive examples designed to show that such beliefs will sometimes be unsafe, so that considerations of safety favour neither side of the debate about KK. Either way, the arguments show that there is no good safety-based reason to reject KK.

7. The Bigger Picture

In sections 3 to 6, we have been concerned with the safety-theoretic objection to KK. But we think that our response really addresses a slightly more general worry.

Regardless of whether one wants to spell out reliability in terms of safety, one might think that the reliability condition on knowledge conflicts with the KK principle. It is natural to think that we can be reliable, while lacking a reliable way of determining that we are; if so, we might know without being able to know that we do.[29] This natural thought can be bolstered by arguing that reliability doesn't iterate: even when an agent reliably avoids error while forming a belief in a scenario, she might not be reliable in those circumstances at reliably avoiding error. It seems to follow that, even when she has reliably avoided error (and thus come to know), she

[29] Cf Alston (1980, pp. 140-141), Williams (1995, p. 96), Antony (2004, p. 12) and Dretske (2004, section 2).

might lack a way of reliably determining that she has reliably avoided error. She will thus lack a way of reliably determining that she knows, and will thus lack a way of coming to know that she knows. Iteration failures for reliability seem to give rise to iteration failures for knowledge.[30]

Our response shows that this more general iteration argument is flawed. Since reliability is basis-relative, and one can form the belief that one is reliable on a basis that entails that reliability, one can form a reliable belief that one is reliable without being reliably reliable. That reliability doesn't iterate thus doesn't show that one can be reliable while lacking a reliable method to determine that one does.[31] And, what is more, we have argued that agents who know do in fact have a reliable method for forming the belief that they are reliable: they can follow KNOW. The more general worry has thus also been defused.[32]

Some might feel that there is something fishy about this response. After all, when it comes to ordinary, first-order beliefs, reliability is a substantive constraint which it's difficult to meet. Why should this constraint be trivialized when the relevant beliefs are about our knowledge?

The answer, we think, appeals not to the details of the arguments just given, but to a core insight of transparency accounts in general: we can base our beliefs about our mental states on those mental states themselves.[33] In our version, reasoning in line with a rule like BEL, allows me to base my belief that I believe that *p* on that belief itself; there is thus no risk of my going wrong, and nothing substantive is required to make my belief safe. There is, admittedly, some question of which mental states can function as a basis for such transparent beliefs: when I reason with KNOW from my knowledge that *p,* am I basing my belief that I know on my

[30] For this argument, see Williamson (2000, pp. 124-127).

[31] This highlights that our reconciliation of reliability and KK differs importantly from that of Greco (2014) and Stalnaker (2015). What they show is that there is a way of understanding reliability (in terms of normal conditions) so that reliability does iterate; we argue that even if reliability doesn't iterate, knowledge still can.

[32] Of course, agents may lack a reliable method for determining *whether* they know, that is, a method that will lead them to believe that they know in (the nearest) worlds in which they do and believe that they don't know in (the nearest) worlds in which they don't. But those sympathetic to *Closure* have independent reason to deny that we need a reliable procedure for determining whether *p* in order to know that p: that requirement would lead to scepticism.

[33] Among theorists who try to explain the possibility of transparent self-knowledge, Dretske (1994) and Gallois (1996) share this commitment with Byrne. A notable exception is Fernández (2013), who thinks that in cases of transparent self-knowledge, an agent's higher-order belief that she believes that *p* isn't based on her belief that *p*, but rather has the *same basis* as her belief that *p*.

knowledge that *p*, or simply on the belief that *p* which that knowledge constitutes? (If the former, the belief that I know will again be trivially safe; if the latter, it will not.) How one comes down on this question will depend, presumably, on how sympathetic one is in general to formulating psychological explanations in terms of knowledge rather than belief.[34] Our arguments about knowledge-extending deductive inference then serve to bring out some of the independent intuitive pressures towards explaining this related psychological phenomenon in terms of knowledge rather than belief; if they are successful, it is only natural to think that I can also base my belief that I know directly in my knowledge. It is thus not surprising that such a belief would end up trivially safe.

This completes our defence of the claim that beliefs formed by following KNOW are reliable. To show this is, of course, not yet to show that the transparency account entails *Restricted KK*. To defend *that* claim, we would have to either argue that reliable beliefs always amount to knowledge, or else that beliefs formed by following KNOW also meet some additional condition C which, together with the fact that they are reliable, guarantees that they are knowledge. Either project is well beyond the bounds of this paper; but since the supposed unreliability of (some) beliefs formed by following KNOW is, we think, the main reason to doubt that they would amount to knowledge, our arguments do show that the transparency account makes *Restricted KK* much more promising than it would otherwise appear.

The transparency account of introspection thus supports KK. Since beliefs formed by following KNOW are always reliable, the transparency account defuses the primary reason to be sceptical of *Restricted* KK, the thesis that an who is able to apply KNOW to the premise that *p* is in a position to know that she knows that *p* whenever she knows p. Moreover, the transparency account also suggests that *Restricted* KK is, despite the unanalyzed placeholder phrase "able to apply KNOW to the premise that *p*", a substantive principle capturing what is appealing about KK; after all, it is formulated by analogy with *Restricted Closure*, and all but the most dedicated opponents of *Closure* should agree that this is a substantive principle capturing what is appealing about *Closure*. If the transparency account is correct, then the usual reasons to expect

[34] For some discussion of this larger issue, see, for example, Williamson (2000, chapters 2 and 3) and Gibbons (2001).

counterexamples to *Restricted* KK fail; and there is independent reason to believe that counterexamples to KK which *Restricted* KK side-steps simply ignore the interesting issue. In this way, the transparency account of introspection offers important new support to KK enthusiasts.[35]

References

Alston, William P. (1980). Level-Confusions in Epistemology. *Midwest Studies in Philosophy* 5 (1):135-150.

Antony, Louise (2004). A Naturalized Approach to the A Priori. *Philosophical Issues 1*4 (1):1–17.

Benton, Matthew (2013). Dubious Objections from Iterated Conjunctions. *Philosophical Studies* 162: 355-358.

Boyle, Matthew (2011). Transparent Self-Knowledge. *Aristotelian Society Supplementary Volume* 85 (1):223-241.

Byrne, Alex (2005). Introspection. *Philosophical Topics* 33 (1):79-104.

Byrne, Alex (2011). Transparency, Belief, Intention. *Aristotelian Society Supplementary Volume*, 85: 201–221.

Byrne, Alex (2012). Knowing What I See. In Declan Smithies & Daniel Stoljar (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press.

Castaneda, Hector-Neri (1979). Philosophical Method and Direct Awareness of the Self. *Grazer Philosophische Studien* 8:1-58.

Cohen, Stewart (1988). How to be a Fallibilist. *Philosophical Perspectives, Volume 2*: 91–123.

Cohen, Stewart (1998). Contextualist Solutions to Epistemological Problems: Skepticism, Gettier, and the Lottery. *Australasian Journal of Philosophy*, 76(2): 289–306.

Cohen, Stewart & Comesaña, Juan (2013). Williamson on Gettier Cases and Epistemic Logic. *Inquiry* 56 (1):15-29.

[35] We would like to thank audiences at the 2015 Edinburgh Graduate Conference in Epistemology, the 2015 Joint Sessions of the Aristotelian Society and the Mind Association, and a number of workshops at MIT, for numerous helpful comments and criticisms. Special thanks are due to Earl Connee, Mikkel Gerken, Jeremy Goodman, Jennifer Nagel, Kieran Setiya, Jack Spencer, Roger White, two anonymous referees, and, especially, Alex Byrne, for crucial objections, insights, and advice.

DeRose, Keith (1995). Solving the Skeptical Problem. *The Philosophical Review*, 104(1): 1–52.

Dretske, Fred (1994). Introspection. *Proceedings of the Aristotelian Society* 94:263-278.

Dretske, Fred (2004). Externalism and modest contextualism. *Erkenntnis* 61 (2-3):173 - 186.

Dokic, Jérôme & Égré, Paul (2009). Margin for Error and the Transparency of Knowledge. *Synthese* 166 (1):1 - 20.

Evans, Gareth (1982). *The Varieties of Reference*. Oxford University Press.

Feldman, Richard (1981). Fallibilism and Knowing that One Knows. *Philosophical Review* 90 (2):266-282.

Fernández, Jordi (2013). *Transparent Minds: A Study of Self-Knowledge.* Oxford University Press.

Gallois, André (1996). *The World Without, the Mind Within: An Essay on First-Person Authority*. Cambridge University Press.

Gertler, Brie (2011). Self-Knowledge and the Transparency of Belief. In Anthony Hatzimoysis (ed.), *Self-Knowledge*. Oxford University Press.

Goldman, Alvin (2009). Williamson on Knowledge and Evidence. In Patrick Greenough, Duncan Pritchard & Timothy Williamson (eds.), *Williamson on Knowledge*. Oxford University Press, pp. 73-91.

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

Greco, Daniel (2015a). Iteration and Fragmentation. *Philosophy and Phenomenological Research* 91 (1):656-673.

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

Hintikka, Jaakko (1962). *Knowledge and Belief.* Cornell University Press.

Lewis, David (1996). Elusive Knowledge. *Australasian Journal of Philosophy*, 74(4): 549–567.

List, Christian & Menzies, Peter (2009). Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy* 106 (9):475-502.

Marušić, Berislav (2013). The Self-Knowledge Gambit. *Synthese* 190 (12):1977-1999.

McHugh, Conor (2010). Self-knowledge and the KK principle. *Synthese* 173 (3):231 - 257.

Moran, Richard A. (2001). *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton University Press.

Nagel, Jennifer (2011). The Psychological Basis of the Harman-Vogel Paradox. *Philosophers' Imprint* 11 (5):1-28.

Neta, Ram (2002). S knows that P. *Noûs*, 36(4): 663–681.

Nozick, Robert (1981). *Philosophical Explanations.* Harvard University Press.

Owens, David (1992). *Causes and Coincidences*. Cambridge University Press.

Pritchard, Duncan (2005). *Epistemic Luck*. Clarendon Press.

Rieber, Steven (1998). Skepticism and Contrastive Explanation. *Noûs*, 32(2): 189–204.

Sorabji, Richard (1980). *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory*. Chicago: University of Chicago Press.

Sosa, Ernest (1999). How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics* 26 (1/2):373-384.

Stalnaker, Robert (2015). Luminosity and the KK Thesis. In Sanford Goldberg (ed.), *Externalism, Self-Knowledge and Skepticism*: *New Essays*. Cambridge University Press.

Strevens, Michael (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Weslake, Brad (2013). Proportionality, Contrast and Explanation. *Australasian Journal of Philosophy* 91 (4):785-797.

Williams, Michael (1991). *Unnatural Doubts: Epistemological Realism and the Basis of Scepticism*. Basil Blackwell..

Williams, Bernard (1978). *Descartes: The Project of Pure Enquiry*. Penguin.

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

Williamson, Timothy (2009a). Probability and Danger. *The Amherst Lecture in Philosophy* 4: 1-35.

Williamson, Timothy (2009b). Reply to Goldman. In Duncan Pritchard & Patrick Greenough (eds.), *Williamson on Knowledge*. Oxford University Press.

Yablo, Stephen (1992). Cause and Essence. *Synthese* 93 (3):403 - 449.