*Episteme* (2026), 1–19
doi:10.1017/epi.2026.10108

episteme

ARTICLE

# Knowing That One Believes Without Knowing That One Knows

Maciej Tarnowski

Institute of Philosophy, Jagiellonian University, Poland
Email: maciej.tarnowski@uj.edu.pl

(Received 1 September 2025; revised 28 December 2025; accepted 17 February 2026)

**Abstract**

In recent epistemology, introspection principles are commonly rejected. One of the central reasons for this is the adoption of Williamson's anti-luminosity arguments (1996, 2000) and the popularity of the associated epistemic externalist position. This rejection, however, comes with theoretical costs concerning the applications of introspection principles in epistemic and doxastic logic and modeling cooperative behavior. In this paper, I provide a way to solve this dilemma by arguing that the principle *KB* – expressing one's privileged knowledge of their beliefs – remains unscathed by Williamson's argument while saving the important theoretical applications introspective principles are used for. I propose a way of justifying *KB* and rejecting *KK* on principled grounds using Byrne's (Byrne, A. (2005). 'Introspection.' *Philosophical Topics* 33(1), 79–104., Byrne, A. (2018). *Transparency and Self-Knowledge*. Oxford: Oxford University Press.) transparency account of introspection, improving upon a previous attempt by Das and Salow (Das, N. and Salow, B. (2018). 'Transparency and the KK Principle.' *Noûs* 52(1), 3–23.). This defense of *KB,* unlike many in the literature, is consistent with epistemic externalism and allows one to reject the problematic *KK* principle and maintain that non-introspective knowledge is guided by Williamson's margin-for-error principle.

**Keywords:** *KK*; transparency; introspection; knowledge; belief; margin-for-error

## 1. Introduction

The epistemic and doxastic positive introspection principles[1], guiding lower to higher-level interaction between epistemic and doxastic modal operators, such as "*a* knows that" ($K_a$) or "*a* believes that" ($B_a$):

[1]While *KK* and *BB* are those most commonly called "introspection principles" for knowledge and belief, to avoid confusion it is easier to address them specifically as "iteration principles" (after Greco 2015a), saving the "introspection" label for a broader application. *KB* (also labeled "positive introspection" by Stalnaker [2006, p. 179]) may be thought of as a plausible strengthening of *BB* since the truth of the latter principle is presumed to be grounded in the agent's infallible capacity for self-*knowledge*, obtained by introspective abilities; assuming that knowledge entails belief, *KB* entails *BB*. *BK* also entails *BB,* though,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

*(KK)* $K_ap \rightarrow K_aK_ap$ *(If a knows that p, then a knows that a knows that p).*
*(BK)* $B_ap \rightarrow B_aK_ap$ *(If a believes that p, then a believes that a knows that p).*
*(KB)* $B_ap \rightarrow K_aB_ap$ *(If a believes that p, then a knows that a believes that p).*
*(BB)* $B_ap \rightarrow B_aB_ap$ *(If a believes that p, then a believes that a believes that p).*

have seen better times. Once widely assumed by philosophers and logicians across the board, proving useful in a variety of different formal endeavors, such as game theory and explanation of strategic behavior (Harsanyi 1967), doxastic and epistemic logic (Hintikka 1962), dynamic pragmatics (Stalnaker 1984), or artificial intelligence (Fagin et al. 2004), now are almost as commonly "rejected by epistemologists of almost every stripe" (Antony 2004, p. 12) as implausible and are, at best, taken to be true only of highly idealized epistemic agents. While sociological reasons for this shift are complicated, it seems that its most profound source is the rising popularity of epistemic externalism, which is often taken to entail their falsity, following the prominent critique of the *KK* principle due to Timothy Williamson (1996, 2000). But should all introspection principles deserve equal treatment or are there important differences between them that may save some of them from externalist critique?

Even though *KK* remains the main point of externalist attacks, all the above-mentioned introspection principles are often assumed to be susceptible to similar pressure, the underlying logic of which is best represented by Williamson's arguments based on the margin-for-error principles. In the next section, I shall briefly present two distinct arguments used by Williamson, utilizing either the unrestricted application of the margin-for-error principle or the agent's knowledge of this principle. I will argue that while the latter argument should be considered sound if one's preferred externalist analysis of first-order knowledge aligns with Williamson's, the former may be still principally resisted from an externalist standpoint. As a result, a possibility for a more moderate view will open – upholding *KB* and *BB* principles and rejecting *KK* and *BK*. In the subsequent sections, I will present an externalist version of a view that combines support for *KB* with the rejection of *KK*, derived from Alex Byrne's (2005, 2018) inferentialist account of self-knowledge. In contrast with the previous analogous attempt by Das and Salow (2018), I shall argue that the most plausible version of Byrne's theory does not commit one to *KK* nor *BK*. I will also discuss how *KB* may secure the most important practical applications of introspection principles, which are often brought up as a form of abductive justification for *KK*. The emerging picture, I shall argue, on the one hand, does justice to popular externalist intuitions regarding the possibility of *KK* failures resulting from "margin-for-error" considerations, but on the other saves the important theoretical applications of introspection principles, as well as the intuition that introspection, unlike perception, testimony or memory, is an especially secure source of knowledge.

Before I start, a few caveats. Firstly, I take introspection principles are not usually taken to describe the actual *performance* of epistemic agents, but rather their epistemic *competence*. For this reason, the higher-order attitude on the right side to the entailment sign is often worded in the literature with explicitly modal phrases, such as "being in a position to know" or "being in a position to rationally believe." I shall, for the most part, ignore these complications here. Accordingly, I will also assume that the agents obeying

apart from merely stating that one has doxastic access to one's beliefs, it also expresses the intuition that from a first-person perspective, believing "feels just like knowing" (Greco 2015b, p. 180). Therefore, this principle is labeled by Stalnaker as "strong belief" and is also accepted in a qualified form by Greco (2014).

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

these principles possess all relevant concepts, such as the concept of "knowledge" and "belief" – that is, they are neither animals, infants nor unsophisticated artificial agents, but adult humans capable of rational reflection[2]. Lastly, in this paper, I shall explicitly focus on attempts to justify the principles in question by appealing to the mechanisms of obtaining self-knowledge by introspection. While most common, this appeal is not universal among the defenders of such principles. For example, Hintikka (1962, pp. 53–59), who accepts only *BB* and *KK*, argues against using introspection in justifying axioms of doxastic or epistemic logic, taking *KK* and *BB* to express a conceptual truth about "belief" and "knowledge." While important, these accounts shall be left out of the scope of the present article.

## 2. Two Williamsonian arguments against introspection principles

As mentioned in the introduction, the most plausible source of contemporary distaste for introspection principles may be traced to the rising popularity of epistemic externalist intuitions and the sympathy such intuitions lend to the margin-for-error principle for knowledge – *MAR* – formulated by Timothy Williamson. Williamson's arguments against the introspection principles universally employ *MAR* and, in turn, his opposition towards them is often monolithically derived from support for the said principle. Yet throughout his writings, Williamson offers at least two logically distinct arguments varying in their scope: the general argument against luminosity of the mental, utilizing the unrestricted form of the margin-for-error principle (as employed in [Williamson 1996] and chapter 4. of *Knowledge and Its Limits* [2000]) and a more specific anti-*KK* argument based the agent's knowledge of this principle (as presented canonically in [2000]'s chapter 5.).[3] In what follows, I shall demonstrate why the epistemic externalist should, in principle, find the latter argument convincing while suspending their support for the former.

Let us first understand why the margin-for-error principle (MAR) may be at all compelling for the broad externalist audience, not necessarily aligned with Williamson's "knowledge-first" approach to epistemology. The principle in question is usually formulated as follows:

(*MAR*) If in case $\alpha_i$ *a* knows that *p*, then *p* is true in case $\alpha_{i+1.}$

where $\alpha_{i+1}$ is a case sufficiently similar to $\alpha_i$. Such cases may be understood as subsequent points of time (e.g., consecutive seconds or milliseconds), situations differing in intensities of a given quality (e.g., minimal-degree increases in temperature), or the size of a perceived object (e.g., 1-inch differences in height of a perceived tree); crucially, the difference between any case $\alpha_i$ and $\alpha_{i+1}$ may be as small as one pleases, up to the point of absolute imperceptibility. The intuitive sense of *MAR* is that for a specific belief to count as knowledge, it need not only be *true*, but *safely true*, that is true in all cases which unrecognizably differ from ours. In this sense, *MAR* can be seen as a generalization of

[2]For a standard objection against an unrestricted application of introspection principles see Castañeda 1970.

[3]While in 2000, Williamson takes his anti-KK argument to rely on assumptions already "implicit in the argument against luminosity" (Williamson 2000, p. 17), he later tacitly acknowledges that *KB* and *KK* need not be offered a uniform treatment (Williamson 2014, pp. 985-989) although he maintains that both are implausible (Williamson 2014, p. 989). The two arguments discussed here are distinguished, e.g., by Ramachandran 2012.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

the factivity condition for knowledge, which covers similar cases, and expresses (with important caveats) the idea that knowledge is supposed to "track truth."

Williamson's first formulations of the principle are used to motivate the epistemicist semantics for vague predicates; later on, he came to hold that much (or all) of our knowledge is *inexact,* and to all such knowledge states, according to Williamson, *MAR* applies. Here is how he introduces the concept of inexact knowledge in *Vagueness*:

> "The notion ( . . . ) is best introduced by examples. Vision gives knowledge about the height of a tree, hearing about the loudness of a noise, touch about the temperature of a surface, smell about the age of an egg, taste about the constituents of a drink. Memory gives knowledge about the length of a walk, and testimony about the physical characteristics of a criminal. The list could of course be continued indefinitely. In each case, the knowledge is inexact. One sees roughly but not exactly how many books a room contains, for example: it is certainly more than two hundred and less than twenty thousand, but one does not know the exact number. Yet there need be no relevant vagueness in the number. The inexactness was in the knowledge, not in the object about which it was acquired." (Williamson 1994, pp. 216–217).

To see how *MAR* applies in those cases given most plausible epistemic externalist views, consider the library case mentioned in this quote. If I am trying to tell how many books there are in a library room by looking, and my eyesight is not perfectly accurate, my belief that there are fewer than 1000 of them, when there are in fact exactly 999, would constitute merely "misplaced confidence," not knowledge. In Williamson's analysis, this is because my belief-forming mechanism (perception) is not sufficiently well correlated with the actual number of books for me to tell the difference between what is actually the case ($\alpha_{999}$) and the case in which there is one more book present ($\alpha_{1000}$). If, on the other hand, I form a belief that there are fewer than 2000 books in the room, this belief counts as knowledge, for it is true in similar cases, while false only in those in which the number of books is 2000 and more, which I would plausibly perceptively distinguish from the actual one and would not form such a belief. These sufficiently similar cases in which *p* remains true form, in Williamson's words, a safety "buffer zone" (2000, p. 19) between the epistemic agent's cognitive success and failure. In this sense, *MAR* explicates a condition of "safety" of knowledge and ensures its "reliability," as present in the traditional externalist analyses of the concept and may be taken to be their formal expression. Hence, even if one believes that knowledge can be given an externalist analysis along the lines of being "safe," "reliably formed" or "truth-tracking" belief, one may be tempted to see *MAR* as approximating their preferred view.

Let us now look at the first, more general argument for the thesis that we may label General Anti-Luminosity. According to this thesis, no non-trivial[4] condition *C* is *luminous* (*L*), in the following sense:

> (*L*) For every case α, if in α *a* is in *C*, then in α *a* is in a position to know that *a* is in *C*.

[4]In Williamson's terms, a trivial condition is one in "which one is always or never in" (Williamson 2000, p. 12).

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

Accordingly, if condition *C* is "knowing that *p*," *L* is equivalent to *KK*, and if *C* is "believing that *p*," *L* is equivalent to *KB* – General Anti-Luminosity entails the rejection of both principles. To set up the argument, we begin by constructing a sorites sequence of ever-so-slightly changing worldly conditions, modeled as a set of cases $\alpha_0, \ldots, \alpha_n$. If we assume that in $\alpha_0$ one is in a luminous condition *C*, we then proceed by utilizing *L* and *MAR* to take us to the conclusion that one is in *C* in any case in the series. To construct the argument against *KK* along the lines of this general strategy, we need the following schema derived from *MAR* for knowledge (*K-MAR*):

(K-MAR) If in $\alpha_i$ *a* knows that *a knows that p*, then *a knows that p* in $\alpha_{i+1.}$

and against *KB,* following schema for belief (*B-MAR*):

(B-MAR) If in $\alpha_i$ *a* knows that *a believes that p*, then *a believes that p* in $\alpha_{i+1.}$

If *K-MAR* and *B-MAR* are true, then knowledge and belief are non-luminous, and hence, *KK* and *KB* are false, by the following argument. Following a classic example, let us imagine Mr. Magoo looking at the modernist clock with just one hand and no numbers and dashes, and no way of knowing the time besides observing the face of the clock from a distance. By looking at the position of the hand, Mr. Magoo, despite his poor eyesight, nevertheless comes to believe something, e.g., that the time is between 3 and 5, or that it is *not* 6 o'clock; those beliefs are accurate and intuitively amount to knowledge (suppose it is 4 o'clock). Let the sequence of cases $\alpha_0, \ldots, \alpha_{120}$ be the sequence of imperceptibly different positions of a hand from 4 to 6; for simplicity, take the difference between any position to be 1 minute. Given that Magoo knows and believes that it is not 6 o'clock given its current position ($\alpha_0$), and, by *KK* and *KB* – he knows that he does know and believe that. By K-MAR and B-MAR, he would then also believe and know that it is not 6 o'clock if the hand of the clock was 1 minute closer to 6 ($\alpha_1$); and here, as well, he would be in a position to know that he does. By applying the same reasoning 120 times, we arrive at the soritical conclusion: Magoo comes to know and believe that it is not 6 o'clock in $\alpha_{120,}$ that is when the position of the hand shows precisely 6 – which is, of course, a false conclusion given that Magoo's perceptual abilities are roughly accurate. Since this result is unacceptable and the fact that Magoo knows in $\alpha_0$ that it is not 6 o'clock is indisputable, *MAR* excludes the possibility of *KK* or *KB* being true. Neither knowledge nor belief is luminous.

The result Williamson arrives at is very strong, given that it generalizes not only over beliefs but also over other mental states that are more commonly thought to be intimately connected with one's awareness – most prominently phenomenal mental states sometimes labeled "self-presenting" (Gertler 2012), such as pains or tickles[5]. What may seem initially suspicious about Williamson's argument is that it derives a highly general and strong conclusion concerning our introspective abilities without even mentioning what sort of process governs the acquisition of introspective knowledge. While intuitive examples used to justify *MAR* usually appeal (as I did two paragraphs above) to our limited abilities of perceptual discrimination, blurry memory, or impreciseness of someone's testimony, making a similar case for introspection proves at least tricky. *However we gain it*, introspective knowledge is accompanied by a strength of authority

[5]Some externalist objections to General Anti-Luminosity limit their scope to defending the luminosity of only these states, e.g., Weatherson 2004 and Conee 2005.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

unseen among other species of knowledge. In other words, if any knowledge is not inexact, the knowledge *that* one believes, perceives, or remembers something to be the case seems to be a good candidate, unlike knowledge *of* the believed, perceived, or remembered fact. Shouldn't introspection be exempt from *MAR* until proven guilty?

Even those generally sympathetic to externalist intuitions have taken issue with this lack of distinction between different sources of knowledge and proposed that *MAR* should be either content-constrained – and not applied to self-presenting states of mind (Weatherson 2004, Conee 2005) – or applied only to certain source-indexed knowledge operators, like *knowledge-by-perception*, *knowledge-by-memory*, and so on (e.g., Dokic, Égré 2009). These proposals however would plausibly not save any of the introspection principles in question. While the former proposal's justification would not typically cover knowledge or belief among self-presenting states, the latter proposal is problematic since it severely limits our ability to model the interplay between knowledge coming from different sources (e.g., gaining *knowledge-by-inference* from applying an inferential rule to premises known by perception and memory). But provisionally exempting self-knowledge from *MAR* does not require such radical moves. To make introspective knowledge immune to *MAR* without rejecting the principle altogether, we may simply assume that the syntactic form of $p$ in *MAR* is limited to that of a sentence free from epistemic or doxastic agent-specific operators (*Restricted MAR*)[6]:

(*Restricted MAR*) If $p$ is a sentence that does not include doxastic or epistemic operators "$B_a$" or "$K_a$," then: If in case $\alpha_i$ $a$ knows that $p$, then $p$ is true in case $\alpha_{i+1}$.

If we adopt the restricted reading, then *K-MAR* nor *B-MAR* no more straightforwardly follow as special cases of the general principle. What then happens to *KK* and *KB*? To answer this question, let us turn our attention to the second version of Williamson's anti-*KK* argument (as featured in 2000, chapter 5.). Unlike in the first argument, let us now stipulate that although Mr. Magoo's higher-order knowledge is not sensitive to *MAR*, he also *knows* that such an amended version of *MAR* is true of his first-order knowledge. If Mr. Magoo knows that he knows that it is not 6 o'clock in $\alpha_i$, by his knowledge of *MAR,* he may still come to know that it is not 6 o'clock in $\alpha_{i+1}$, for he knows that whenever he knows that it is not 6 o'clock in $\alpha_k$ for some $k$, it remains true in $\alpha_{k+1}$. This, in turn, validates *K-MAR* as applied to knowledge of one's first-order knowledge (assuming additionally one-premise closure) without assuming anything of substance about one's introspective abilities – and we may again derive a paradoxical conclusion. In effect, to reject *KK* via the second route, we trade the unrestricted-order application of *MAR* for an additional assumption concerning one's knowledge of *MAR*.

How far does this argument extend? Quite intuitively, one may still apply it to *BK*, essentially replicating the above reasoning under the scope of belief operator. If Mr. Magoo believes that it is not 6 o'clock in $\alpha_i$, by *BK* he also believes that he knows that; and if he believes even in our restricted, first-order *MAR,* he still may come to believe that it is not 6 o'clock in $\alpha_{i+1}$ by *modus ponens* reasoning; and so, the argument further

[6]This crude limitation comes with some more or less obvious caveats. Most importantly, the language used needs to disallow the possibility of encoding higher-order beliefs into first-order ones; it will be sufficient for our purposes to take our model to be that of formulas of propositional multimodal logic. We also abstract here from instances of *testimonial* or *perceptual* self-knowledge that cannot be otherwise obtained by introspection.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

replicates. The same does not, however, easily extend to *KB*. If one rejects the idea expressed by *BK* – that we are committed to treating our beliefs *as knowledge states* – then one has no good reason to believe that the restricted *MAR* principle holds of one's belief; they may deny that they would still believe that *p* is true in $\alpha_{i+1}$ even if they claim to know that they believe *p* in $\alpha_i$. Therefore, accepting the latter, but not the former Williamsonian argument allows for accepting *KB* (and *BB* as a result), but not *BK* and *KK*. Let me label this combination of views ∼*KK*+*KB* thesis.

As shown above, upholding ∼*KK*+*KB* without completely abandoning *MAR,* but just restricting it, is at least consistent. Whether it is a *good* position to occupy comes down to answering whether self-knowledge of beliefs is sufficiently exact to justify the mentioned restriction. In line with the idea that the truth of these principles ought to be derived from a characterization of our introspective abilities, answering the above questions requires rooting in a sufficiently strong epistemological account of self-knowledge. To fulfill our purpose, such a theory should both (a) vindicate the idea that higher-order knowledge of one's beliefs is not subject to *MAR* and at the same time (b) stay faithful to the externalist intuitions about knowledge motivating restricted *MAR*'s application. The following sections will discuss the prospects, limitations, and challenges of exactly such a project.

## 3. Transparency of belief and *KB*

In *The Principles of Psychology*, William James pronounces that "introspective observation is what we have to rely on first and foremost and always. The word introspection need hardly be defined; it means, of course, the looking into our own minds and reporting what we there discover." (James 1890, p. 185). While his epistemic optimism towards the infallible reliability of introspection is (as witnessed above) not as popular today, his vision of the source of self-knowledge is more prevalent and widely represented. Starting from Locke and Hume, many philosophers characterized introspection as an "inner sense" or "inward perception," and knowledge of one's mental states as essentially perceptual – most notable contemporary examples include Armstrong (1968) or Nichols and Stich (2003). This vision of introspection also seems to be the one endorsed by Williamson, e.g., when he justifies the intuitive rejection of *KB* by noting that it is "at least as hard to introspect the time according to [one's] phenomenal clock as to see the time according to a real clock" (2014, p. 989), presupposing that we come to know what we believe by *observing* how things appear to us – presumably, with a "mind's eye." It is easy to see why this account of introspection makes the application of *MAR* to self-knowledge compelling. For if knowing that one is in mental state *M* comes from one *perceiving M*, this perception plausibly must admit at least *some* possibility of error. As such concession would validate *MAR* when applied to introspective knowledge, I believe this approach to introspection should be abandoned, if we are set on justifying *KB*.

Among competitors to "inner sense" theories, we might crudely distinguish at least two of their bundles: *acquaintance theories*, building upon Bertrand Russell's remarks on acquaintance with the mental, and *transparency theories*, building on Gareth Evans' work (1982).[7] As the former's stronghold field of application is knowledge of one's phenomenal states (as, e.g., Gertler 2012, pp. 102–104 directly concedes[8]) and the

[7]For a more fine-grained taxonomy, see Gertler 2021.

[8]Russell (1912, pp. 76-80) originally claimed *all* introspective knowledge to be based on acquaintance, but such a strong view is rejected by most contemporary acquaintance theorists, such as Brie Gertler or Laurence BonJour, along with much of foundationalist epistemology endorsed by Russell in *Problems* . . . .

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

latter's – knowledge of one's epistemic states, it is natural to focus on transparency theories when arguing for the luminosity of belief expressed by *KB*. According to transparency accounts of self-knowledge, the agent's knowledge of their beliefs[9] is not acquired by anything close to perception of our "inside world," but rather by attending directly to the content of such beliefs. In the often-quoted words of Gareth Evans, "in making a self-ascription of belief, one's eyes are, so to speak, (. . . ) directed outward – upon the world": to answer the question of whether one believes that *p*, one needs to attend to "precisely the same outward phenomena" as one would when answering the question whether *p* is true (1982, p. 225). Call this particular process of figuring out what we believe *the transparency procedure*.

Can utilizing the transparency procedure for one's beliefs yield knowledge? And, if it can, what *kind* of knowledge is generated by following this procedure? There are many responses to these two questions present in the literature. Since the purpose of this paper is not to adjudicate between them[10], I shall follow one of the most detailed and popular transparency accounts of introspection proposed by Alex Byrne (2005, 2018). As we shall see, Byrne's account is especially promising for our project of finding the theory fulfilling the starting desiderata due to its simplicity, wide scope, and straightforward compatibility with externalist epistemology.

In short, Byrne's approach to the problem could be summed up as describing the transparency procedure straight-up as a form of *inference*, thereby grounding our capacity for self-knowledge in a naturalistic competence non-controversially possessed and mastered by human agents. More specifically, such transparent inference is described as utilizing a specific set of *epistemic rules*, i.e., rules explicitly guiding one's acquisition of belief expressed by a conditional of the form R (Byrne 2018, p. 101):

R: If conditions *C* obtain, believe that *p*.

Inspired by an earlier proposal of Gallois (1996), Byrne proposes to describe the agent acquiring transparent self-knowledge of beliefs as an agent trying to follow a BEL rule of the form:

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

Obtaining higher-order beliefs with the use of BEL is described as follows. On Byrne's account (2018, pp. 101–102), one *follows a rule* if one *knows* the premises of the rule and *tries to follow a rule* if one merely (and perhaps erroneously) *believes* them. To be in a position to apply BEL, i.e., follow or try to follow BEL, one then has to either know or believe that *p*. Hence one's higher-order belief resulting from applying BEL – that *one believes that p* – will come out true whenever one is in a position to apply BEL. By this process, one therefore is guaranteed to have accurate higher-order beliefs about their mental states. Moreover, it grounds our introspective process in inferential abilities which are uncontroversially possessed by ordinary agents and the success of applying the rule is properly limited to the first-person perspective. This, in turn, allows Byrne to

[9]Many authors disagree about the exact scope of the transparency procedure – whether it applies only to beliefs (as e.g., Moran 2001 seems to suggest), or to other mental states as well. Byrne (2018) maintains that *most* of our introspective knowledge is generated by a transparency-like procedure: I shall restrict myself here to considerations of belief and knowledge.

[10]Burge (1996), Moran (2001), and Boyle (2011) are other detailed accounts.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

account for both "peculiar" and "privileged" access to one's beliefs (see Byrne 2018, pp. 108–112), i.e., the intuition that one knows that they have certain beliefs *differently* than others know it of them, and enjoys epistemic *authority* concerning this fact.

Can higher-order beliefs acquired by applying BEL be, however, adequately described as knowledge states? Many of the criticisms pointed against Byrne's account (e.g., Boyle 2011) mention the fact that the inference supposedly performed when one applies BEL is very much unlike any other type of inference that is supposed to yield knowledge. There is neither a causal nor evidential connection between the truth of $p$ supposed in the premise and the fact that one believes it, stated in the conditional's conclusion; $p$ does not make the fact that *one believes that p* anyhow more likely. Moreover, the explanation of BEL's success explicitly allows one to acquire knowledge by reasoning through a false step, for $p$ might very well be false while the true conclusion is to be granted the status of knowledge. In short, to use Boyle's phrase, "the inference is mad" (2011, p. 230) and does not fit any of the ways standardly associated with justifying an inferential rule.

While Byrne developed detailed responses to such counterarguments (2018, pp. 121–127), their crucial point, which closely aligns his position with the externalist intuitions we seek, is emphasizing the unmatched *reliability* of BEL-based inference as a belief-producing mechanism.[11] The mere honest application of the BEL rule, unlike other rules of inference, secures the truth of the resulting belief irrespective of whether one actually follows or merely tries to follow BEL – in Byrne's terms, this makes BEL *strongly self-verifying*. One may usefully compare BEL with another epistemic rule:

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

or a *modus ponens* inference couched in terms of an epistemic rule:

CLOSURE: If $p$ and $p$ *implies* $q$, believe that $q$.

While both rules are knowledge-generating, neither of them produces *only* true beliefs if applied. For example, unlike BEL, CLOSURE is not strongly self-verifying, for one's false belief in the conditional or its premise may lead one to a false conclusion. In Byrne's terminology, CLOSURE counts as merely self-verifying, as it yields knowledge only if one follows, not just tries to follow it. The rule like DOORBELL is not even that – one's knowledge that the doorbell rings does not thereby guarantee the presence of someone at the door. Though arguably both CLOSURE and DOORBELL reliably produce true beliefs, it would be superficial at best to ground their validity in this fact only, as their reliability as epistemic rules is itself grounded in the validity of a logical rule of inference or a pattern of "best explanation" reasoning. A brain in a vat (or a counterpart of ours deceived by an evil demon) would be blameless in trying to follow CLOSURE and DOORBELL even though beliefs produced by applying them would turn out to be at large false, but they would be blameless only because these rules are fashioned after otherwise justified patterns of inference. On the other hand, a brain in a vat trying to follow BEL would still end up with true higher-order beliefs about their beliefs and could be advised to apply BEL on these grounds alone. One may require the above traits – "no

[11]Though Byrne does not endorse a reductive analysis of knowledge in terms of the reliability or safety of beliefs, he does endorse a broadly externalist epistemology that perceives these features as essential properties of beliefs deemed to be knowledge states (Byrne 2018, pp. 103-109).

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

false lemmas," the logical or evidential connection between the premise and the conclusion – from an epistemic rule insofar as the rule is *not* strongly self-verifying and the risk of error arises. If a rule *is* strongly self-verifying, this feature alone seems to justify its use on brute, reliabilist grounds.

Does our reliable BEL in some way grounds the validity of the *KB* principle and allows for exempting knowledge gained through it from the scope of *MAR*? To answer this, let me first propose an adjusted version of *KB* (*Adjusted KB*) rooted in the assumption that BEL is a proper epistemic rule:

(*Adjusted KB*) If *a* believes that *p* and *a* is able to apply the BEL rule to the premise that *p*, then *a* is in a position to know by applying BEL that they believe that *p*.

The adjustment (modeled after Das and Salow's [2018, p. 8] modification of *KK* that shall be discussed in the next section) counters the standard objections to introspection principles, such as a purported lack of the concept of belief necessary to *form* the higher-order belief at all or the lack of computational power necessary to perform relevant deductions. On the other hand, however, this modification makes more vivid the similarity between *KB* and other principles of doxastic and epistemic logic that may be grounded in epistemic rules as described by Byrne. If the above-mentioned standard objections are sound, very similar ones could be raised, e.g., against epistemic or doxastic closure principles. Surely, an infant might not be able to apply CLOSURE to the premises *p* and *p implies q* because of their conceptual or computational limitations; one might also be too tired or distracted to exercise their usually possessed ability. What matters in these considerations is the agent's *competence*, not (the limits of) their *performance*. Byrne's theory, by positing that knowledge of one's beliefs is rooted in one's inferential, not perceptual, abilities, then allows one to analogize between *KB* and other less controversial principles of epistemic-doxastic logic, such as epistemic or doxastic closure. Its crucial prediction is that once one believes that *p*, then one knows *all there is to know to know that they believe it*, just as once one knows *p* and that *p implies q*, then one knows all there is to know to know that *q*, given they have relevant inferential capacities.[12]

To decisively answer whether such justification is enough to secure the needed *MAR* exemption, one may still consider the following modification of Williamson's argument sometimes directed specifically against *KB*. Let us imagine the case, in which one's confidence in *p* very gradually declines over time $t_0 - t_n$, so that in the moment $t_0$ one believes that *p*, but does not at $t_n$. It is independently plausible that (however we gain knowledge of our credence) our ability to determine our level of confidence is not precise. Hence, on most occasions, one would be unable to discriminate between having

[12]A natural question (though unaddressed by Byrne) here is: are epistemic rules like BEL learned or innate? The above characterization makes it seem like a learned competence, which, in turn, makes it troubling to grant them „self-verifying" status. The question, in my view, is likely to be decided by considering whether the concepts used in these rules, as well as inferential capacities in general, are learned or innate. When it comes to BEL, "belief" may be plausibly characterized as more or less hardwired, given its role in spontaneous metacognitive processes and primitive forms of mindreading, and nothing in the above characterization requires one to apply the BEL rule by *explicitly* representing the rule in question, so that it would have to be learned. To say that the competence to apply BEL is in some way innate is not the same as to say that one is *always* in a position to apply it, nor that such competence cannot be more or less robust at different developmental stages. My thanks go to the Reviewer for this journal for pressing me on that point.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

credence $r$ at $t_i$ and a slowly lower credence $r'$ at $t_{i+1}$. If belief is, as many take it to be, intimately connected with sufficiently high confidence, this seems to lend independent plausibility to *B-MAR,* as *knowing* that one has high confidence in $p$ at $t_i$ seems to require that one maintains high confidence in $p$ at $t_{i+1.}$ To adjust this reasoning to our context, let us now stipulate that the agent in question is able to apply BEL at any of $t_0$ – $t_n$ moments. Isn't it then possible that at some threshold point, one mistakenly applies BEL because of their misplaced confidence in $p$ – which, even though still high, just barely falls short of full belief – and come to falsely believe that they believe that $p$? To put it differently: is it possible that the agent not even tries, but merely *tries to try* to follow BEL? If such cases are possible, then we should reject the *Adjusted KB*.

To see how one could reject this argument on Byrne's account, let's note that this reasoning rests on the premise that it is the agent's having high confidence in $p$, not just their belief that $p$, triggering their disposition to apply the BEL rule. In line with the logic of the argument, there must be both the largest $i$ such that at $t_i$ $a$ believes that $p$, while at $t_{i+1}$ $a$ no longer believes that $p$, as well as the largest $j$ such that at $t_j$ $a$ is disposed to apply BEL to the premise that $p$ but is not disposed to do so at $t_{j+1,}$ and finally the largest $k$ such that at $t_k$ $a$ has (sufficiently) high confidence in $p$, but does not at $t_{k+1}$. For the argument to be sound, it needs to be established that $i > j \geq k$, that is, that one is disposed to still apply BEL once they lose the belief, but retain sufficiently high confidence in $p$. This ordering is however not at all obvious. Note that on Byrne's characterization, to apply the BEL rule to the premise that $p$, one needs first to judge that $p$ simpliciter, not only that $p$ *is likely* or even that $p$ *is highly plausible* (unless one wishes to find out whether they *believe that p is likely*). Therefore, to acquire second-order beliefs through BEL, the agent needs to be disposed to judge that $p$ and use $p$ as a premise in theoretical reasoning (namely – BEL). These dispositions are frequently taken to be important properties of belief states (e.g., Audi 1994, Silins 2012, Smithies 2019, also: Williamson 2000, pp. 46–47); and even if we are not full-on dispositionalists about belief, one may plausibly claim that exhibiting such inferential dispositions is essential, and so $i = j$. Adopting this premise does not entail rejecting the claim that belief is intimately connected with confidence. A fully committed Lockean might hold that belief is equivalent to some subjective credence above some threshold $r$, and grant that on many occasions one is not in a position to know what precise confidence one has in a certain proposition p. Yet until they do not also maintain that at the same time, one is in a position to know what the appropriate threshold is[13], it is consistent to maintain that one may figure out by applying BEL whether one believes $p$, without knowing what precise credence underlies this belief.

If such judgment-generating and inferential dispositions are essential for belief and Byrne's characterization of the transparency procedure is right, it is then quite easy to see how the *B-MAR* principle may fail, and *KB* be true. On Byrne's inferential picture, coming to know that you believe that $p$ requires just *exercising* the relevant dispositions, not knowing their exact boundaries. Hence, even though one cannot know exactly where one's confidence threshold for belief lies on a particular occasion, this fact does not justify the *B-MAR* principle. If $a$ believes that $p$ at $\alpha_i$ but not at $\alpha_{i+1}$, then $a$ is disposed to judge that $p$ and apply BEL to the premise that $p$ – and come to know that they believe that $p$ – only at $\alpha_i$. As one is in a position to use the BEL rule only in success cases, the

[13]Unless one maintains the "credence 1" view of belief (Greco 2015b) – but in this case, it is plausible that any form of doubt stops such an agent from applying BEL.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

threat of error does not arise.[14] Byrne's theory allows us then to account for both of our starting desiderata. It secures one's strong introspective access to one's beliefs as a source of knowledge by appealing to the reliability of the transparency process (thereby validating the *KB* rule) and on the other hand can avoid the margin-for-error considerations concerning the process' reliability, which allows to describe our second-order knowledge generated by BEL as peculiar and distinct from first-order knowledge gained from experience or inference. Let us now see how the above discussion translates to the debate surrounding *KK*.

## 4. Transparency of knowledge and *KK*

One thing I demonstrated in the previous section was that Byrne's transparency account provides an externalist case for *KB*; what needs to be shown is that this case extends also to $\sim KK+KB$. I am not the first one to point out the connection between Byrne's theory and the justification for introspection principles – the first such case was made by Das and Salow (2018).[15] Yet, the focus of their approach was to use the transparency account to defend *KK*. As I wish to do the exact opposite, I need now to rebut their argument. In the course of this rebuttal, I will also propose some independently plausible augmentations to Byrne's transparency account and demonstrate how it may ground an externalist argument against *KK* different from that of Williamson.

In making their case for *KK*, Das and Salow do not make use of BEL but the epistemic rule KNOW, proposed by Byrne (2012):

KNOW: If $p$, believe that you know that $p$.

On their account, if we assume that the agent is in a position to apply the KNOW rule and follows it (in Byrne's sense of rule-following), one's resulting belief will be guaranteed to be true in the same way as the belief produced by BEL. Because unlike BEL, KNOW is not *strongly* self-verifying, i.e., the resulting belief comes out true only if the agent actually follows KNOW and not merely tries to follow it, Das and Salow claim that the belief resulting from the application of a self-verifying epistemic rule qualifies as knowledge *only if* the agent qualifies as following the rule in Byrne's sense. Hence, whenever one comes to believe that one knows that $p$, although one does not know that $p$, their belief is a result of merely *trying to follow KNOW* which, according to Das and Salow, is insufficient to generate knowledge (2018, p. 10). If one, on the other hand, *follows KNOW*, one is epistemically guaranteed to succeed – and *KK* (in a qualified form analogous to *Adjusted KB*) gets validated. Das and Salow support their case for this restriction by likening the knowledge obtained by following Byrne's transparent rules to the knowledge obtained by following rules like CLOSURE (2018, p. 5).[16] Their argument

[14]Can one be then ever *mistaken* about their own beliefs? The answer suggested here is: no, if the second-order belief in question was formed by applying BEL. This does not preclude the possibility of obtaining such mistaken higher order beliefs in some other way, e.g., by testimony or behavioral self-observation. Byrne takes this to be possible, though atypical (Byrne 2005, pp. 83-84; 86-87); I prefer to suspend judgment on this matter.

[15]A similar argument to Das and Salow's, though not explicitly based on Byrne's account, is made in McHugh 2010. Byrne himself seems to remain agnostic as to whether his account provides any support for introspection principles (Byrne 2018, p. 116).

[16]Instead of CLOSURE, they use OR, epistemic rule based on disjunction introduction. The choice of the rule does not matter here.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

is fairly simple: given that CLOSURE, like KNOW, does not generate knowledge if one does not *know* the premises *p* and *p entails q,* one's epistemic base for gaining knowledge through epistemic rules needs to be *following,* not *trying to follow a rule* (2018, pp. 11–12). Because we should count all sufficiently similar rules governing the acquisition of inferential knowledge as having the same type of epistemic base (by what Das and Salow call *Generality Constraint* [2018, p. 11]), it follows that all self-verifying epistemic rules should count as knowledge-generating *only if* the agent follows them, not merely tries to; this, in turn, gives them the right to claim that only beliefs resulting from *following* KNOW may count as knowledge, and validate *KK*.

As we may already see, the case for *KK* Das and Salow proposes differs significantly from the simple reliabilist justification provided above for *Adjusted KB* based on BEL. Since KNOW produces true beliefs only if followed, the question of why it could be said to be knowledge-producing no longer can be answered on the grounds of the exceptional ability to produce only true beliefs. A brain in a vat will no longer easily gain higher-order knowledge by applying KNOW; it is also quite plausible that a purely statistical assessment of reliability will qualify only some – more fortunate or epistemically cautious – actual world agents as forming reliably true beliefs by applying KNOW. Their justification for why KNOW may count as a knowledge-generating rule needs then to be substantially different. BEL's status as a *good* epistemic rule is made by appealing to the fact that it is *strongly self-verifying*, which implies its infallible reliability in producing true beliefs. A similar case cannot be made for KNOW because of its *self-verifying* status, for it is simply not the case that all self-verifying epistemic rules deserve a place in our cognitive repertoire. If we agree with Williamson even on *Restricted MAR*, then the following becomes an example of a self-verifying epistemic rule:

SORITES: If p is true at $\alpha_{i,}$ believe that p is true at $\alpha_{i+1.}$

since knowing that *p is true at* $\alpha_i$ (i.e., *following* SORITES) guarantees that *p is true at* $\alpha_{i+1.}$ Yet, as we may readily see, thoroughly applying SORITES to one's beliefs yields detrimental results in the form of a variety of fallacious soritical conclusions. The crucial difference between SORITES and CLOSURE in this respect is that while CLOSURE is *knowledge-conducive*, SORITES is not. Though if one knows that *p is true at* $\alpha_i$, then their SORITES-based belief that *p is true at* $\alpha_{i+1}$ would be true, there is a point at which it would fail to be knowledge. While KNOW guarantees that, if followed, it produces true beliefs, we have no justification as to why such beliefs should constitute knowledge.

Should we be skeptical of KNOW then? Given the fact that the rule neither shares its reliability with BEL, nor is independently justified (by logical principles, like CLOSURE, or abductive inference, like DOORBELL), I think the answer is yes. A more straightforward reason for such skepticism is that, if accepted, KNOW lends support to the (appropriately adjusted) *BK* principle – which, as we saw above, is inconsistent with *Restricted MAR*. In fact, much of the intuitions supporting KNOW, such as that we are "rationally committed to treating our beliefs as knowledge" or that "belief feels like knowledge," seem to parallel those offered in support of *BK*.[17] If we, however, are on the

[17]That seems true of Das and Salow and others, who find *BK* plausible (e.g., Greco 2014). Byrne himself makes his case for KNOW rather brief, by pointing out an analogy between Evans' transparency procedure for self-ascriptions of belief and self-ascriptions of knowledge:

> "If someone asks me 'Do you know that it's raining?' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Is

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

right track in justifying introspection principles by appealing to Byrne's epistemic rules while acknowledging the plausibility of *MAR* for first-order knowledge, these intuitions need to be treated with caution. For while we might concede that in ordinary cases one cannot internally distinguish between merely believing and knowing that *p*, it does not follow that one always takes one's beliefs to be knowledge states, just that one usually cannot determine whether a certain belief of theirs constitutes knowledge. From an externalist point of view, it is perfectly coherent to hold that we can be agnostic concerning our beliefs constituting knowledge on a particular occasion. Think again of Mr. Magoo observing the clock. When he comes to believe at 5.50 that it is not 6 o'clock, he may still plausibly suspend judgment as to whether his eyesight is good enough to grant him knowledge that it is not 6 o'clock, and yet fully believe that the world is the way it presents itself to him, even if only for the fact that there is no better data he could rely on in his reasoning. If his skepticism towards his perceptual abilities is unwarranted, his belief constitutes knowledge, if it is warranted – then it does not. But while he has no means to assess which of the scenarios he is in, and hence cannot truly distinguish "from within" whether he merely believes or knows that it is not 6 o'clock, it does not seem like a form of irrationality on his part to stick to believing what seems to him true.

What should we then do with KNOW if we reject the intuition of the first-person "knowledge-belief collapse"? I suggest that staying true to externalist intuitions, we may propose the following Byrne-style epistemic rule to guide our self-ascriptions of knowledge:

KNOW*: If you believe that *p* and your belief that *p* has high epistemic credentials, believe that you know that *p*.

KNOW* is a transformed CONFIDENCE rule, proposed by Byrne (2018, pp. 119–121) to guide our self-ascriptions of high (and low) confidence in belief. By "high epistemic credentials" ascribed to the belief I mean, like Byrne, relevant evidence about the source of the belief and reliability of the belief-forming mechanism. Importantly, unlike in KNOW, self-ascriptions of knowledge are here "broadly Rylean," i.e., based on the assessment of the reliability of our cognitive faculties, perception, memory, etc. that is in principle available also to others. Consider, for example, my beliefs that the Warsaw Ghetto Uprising took place in 1943 and that it will rain in 3 days. Even though both are my beliefs, I am more inclined in giving the first, but not the second one the status of

it raining?'. This sounds equally plausible as the original claim about belief, suggesting that I know that I know that it's raining by following [KNOW]" (Byrne 2018, p. 116).

There is a good reason to distrust the force of this specific epistemic-conversational pattern. Why? Because there is an important point of disanalogy between "Do you believe that *p*?" and "Do you know that *p*?" questions: while the first does not involve any non-trivial pragmatic assumptions about the truth of *p*, the second one, in normal circumstances, involves the speaker presupposing that *p* is true (i.e., is knowledge-apt in the first place). If we employ a transparency procedure when answering such questions, this fact could be easily explained away by this pragmatic assumption: what one queries in asking such a question is our state of mind – whether we believe that *p* – not our epistemic warrant or reliability – the speaker is pragmatically expected to know that *p* already. On the other hand, the question "*Do you only believe that, or do you know it?*" is *not* a nonsensical one, and the agent may formulate a sensible answer to it. If I assert that *p*, my interlocutor may direct this question to me. In answering them, it seems that I should attend to phenomena peculiar to the epistemic status of my belief, such as the reliability of my source, instead of reiterating the transparency procedure; the pragmatic emphasis shifts from querying my opinion to interrogating its safety.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

knowledge. Why? In case of the first belief, I may consider a variety of factors that make such status plausible: (a) that someone trustworthy have told me that, (b) that I am generally well-acquainted with World War II history, (c) that I'm an inhabitant of Warsaw, (d) that I am good at remembering dates. All these factors speak to *my reliability* in forming and maintaining the belief in question, so they are directly relevant evidence for my belief constituting knowledge on the externalist picture – they *objectively* make it more likely that my belief constitutes knowledge.[18] On the other hand, I can think of much fewer relevant factors giving high epistemic credentials to my belief about the future weather; even if I remember checking a weather app on my phone, my assessment of the long-term reliability of meteorology is rather poor. Hence, even though this belief *may* constitute knowledge (the testimony of the app may be reliable), it is not subjectively plausible, and hence I would be well advised to suspend judgment on the matter. Similarly, a person who knows all these relevant factors may be in principle just as justified in granting or denying my belief a status of knowledge – and hence I have no *peculiar* or *privileged* access to the epistemic status of my belief.

Given the interpretation of MAR advanced here, it is natural to include safety from error within a reasonable margin as one of the factors that belong to the assessment of the epistemic credentials of our belief. This allows to accommodate the Williamsonian intuition that the imprecision built into our belief-forming mechanisms may prevent us from knowing that we know something even if we do know it. Mr. Magoo may very well come to believe that it is not 6 o'clock by looking at the clock showing half past 5 from a distance, but still, after reflecting on his poor sight and discrimination abilities, refrain from believing that he knows that, as his epistemic credentials do not meet the desired standard. Yet, if this reflection leads him to believe that his perceptual faculties are appropriately fine-tuned and the risk of error is minimized, he may come to believe (and, if he is right – know) that he knows that it is not 6 o'clock.

All this seems quite natural for an epistemic externalist. If an externalist takes themselves to know that knowledge that $p$ requires sufficient margin-for-error, they may be expected to be more conservative in their self-ascriptions of knowledge than that of belief. Knowing that one knows, unlike knowing that one believes, requires also knowing that their first-order belief meets relevant externalist criteria[19], which feature here under the label of "high epistemic credentials." Only when one believes that their belief satisfies these criteria, they may conclude by KNOW* that they know that $p$. Unlike BEL and KNOW, KNOW* is not self-verifying – the protagonist of the Gettier scenario would arrive at a false conclusion that they know that $p$ despite *following* KNOW*. However, in

[18]Byrne makes essentially the same point in (Byrne 2018, p. 120), even though he does not find it inconsistent with KNOW; he merely says that „[e]vidence that one's belief is not knowledge can trickle down to the belief itself, lessening one's confidence in the proposition believed" and thus argues for the relevance of these considerations to one's self-ascriptions of subjective confidence. I find it independently psychologically plausible that similar considerations are relevant in ascriptions of confidence, and that often what we take ourselves to *know* and what we are most confident in coincides.

[19]One may protest that this reasoning commits an intensional fallacy, for according to externalist standards, it does not matter whether the agent's belief that they know that $p$ is supported by such robust evidence, but only that it is properly formed (Okasha 2013). However, if we grant that the agent believes that first-order knowledge requires, e.g., Williamsonian margin for error, then they cannot believe that they know that $p$ without believing that their first-order belief meets *MAR* on pain of doxastic inconsistency (see Bird and Pettigrew 2021, pp. 1726-1727). Hence, since to run Williamson's argument for first-order knowledge we *need* to assume that the agent knows or believes *K-MAR*, we may stipulate that such conditions obtain.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

opposition to these rules and much more like DOORBELL, KNOW* is abductively justified: that one believes that $p$ with high epistemic credentials *is* good evidence for them knowing that $p$. Quite easily we can then classify KNOW* as knowledge-generating, without the troubling consequence of collapsing belief and knowledge from the first-person perspective, upholding skeptical conclusion that we are never in a position to know that we know anything or denying the strong introspective access to one's own beliefs. Hence, the epistemic externalist is offered a way out by modifying Byrne's theory to better reflect their intuitions concerning justified self-ascriptions of knowledge, while retaining its strength with respect to justification of self-ascriptions of belief. If one follows Byrne in taking introspective knowledge that one sees (2012) or remembers (2018, pp. 184–195) as inferential, one may readily apply the above considerations to such cases as well. Given that *seeing* and *remembering* are factive mental states, the rules SEE and MEM Byrne proposes should be also weakened in a manner similar to KNOW*, while their strongly self-verifying counterparts would give one instant knowledge of the fact that one perceives or has a recollection experience.

To finish this section, let me briefly discuss the larger import this discussion has on related issues of application of introspection principles. Why go through all this trouble, you may ask? After all, even if I succeed in defending $\sim KK{+}KB$, the exercise may be just that of coherent expression of a certain class of epistemic intuitions, without any practical or theoretical import. It may seem that giving up on *KK* may be too much for all the relevant practical purposes. Contemporary defenders of *KK* (e.g., Greco 2015a) usually point out its two important applications: the treatment of certain epistemic Moore-like paradoxes and the natural explanation of how groups of collaborating agents might achieve common knowledge. Let me now briefly sketch how upholding $\sim KK{+}KB$ might allow us to mitigate worries connected with losing these applications, and why *KB*, not *KK*, actually might be *the* principle needed for the job.

Even though, obviously, certain results hold only if *KK* is assumed, many similar results hold if we accept *KB* instead. In doxastic logic, the *BB* principle often allows us to prove theorems that are obtainable in epistemic logic under the assumption of the factivity of knowledge (expressed by axiom T), but not the weaker assumption that it is merely consistent (expressed by axiom D). Moore-paradoxical sentences, i.e., sentences of the form $p \,\&\, \sim B_a p$ or $p \,\&\, B_a \sim p$ and similarly puzzling "anti-expertise" sentences of the form $p \leftrightarrow \sim B_a p$ can be proven to be rationally unbelievable for any *KD4*-doxastic agent $a$, while their epistemic analogs – $p \,\&\, \sim K_a p$ and $p \leftrightarrow \sim K_a p$ – are unknowable simply for epistemic *KT*-agents.[20] Therefore, if one wishes to maintain that there is a significant analogy between the unknowability of certain epistemic Moorean constructions and the unbelievability of their doxastic counterparts, one may be advised to accept *KB* (and *BB* as a result). As for common knowledge (spelled out in terms of an infinite hierarchy of mutual knowledge about epistemic states of others), it is quite unclear in the first place whether this assumption needs to target *knowledge* as a first-order mental state. We may, and often do, rationally coordinate even if we all believe something false; whether our first-order mental state is a *factive* state usually does not explain our success in communication, collaboration, or competition. In this vein, many philosophers write of "knowledge"-part of "common knowledge" as a concept "which can accommodate the medieval 'common knowledge' that the Sun circles the Earth"

[20]While belief in standard Moorean conjunctions can be avoided if we strengthen the *KD*-logic of belief with the "do not believe what you believe you don't know" axiom ($Bp \rightarrow \sim B{\sim}Kp$), the same does not hold for "anti-expertise" doxastic sentences, as they are *KT-knowable*. I write more on this in Tarnowski 2025.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

(Heal 1978, p. 116) and treat the assumptions of common knowledge and common belief as essentially interchangeable for all intents and purposes (Lederman 2018). For the latter, but not the former, the assumption of *BB* is all that is required. While there has been some doubt towards such interchangeability (see e.g., Yalcin 2024), this skepticism typically targets the fact that *common belief* allows for false or Gettiered higher-order beliefs, which disables coordinational success or makes it a matter of pure luck. Opting for stronger *KB* escapes this problem since in the standard way it secures common *knowledge* of first-order belief if agents assume that their knowledge is symmetric[21], as in normal doxastic-epistemic logic *KB* entails the following principle:

$(K^nB)$ $B_ap \rightarrow K_a{}^nB_ap$ *(If a believes that p, then a knows*$^n$ *that a believes that p).*

where "knows$^n$" means "one knows that" repeated *n*-times for any *n*. Thereby *KB* disallows Gettiered higher-order beliefs and responds to this objection, at the same time accommodating the intuition that a group of agents may rationally coordinate or compete given they all commonly know what all agents *believe* even if such beliefs are false or unjustified.

## 5. Conclusion

The overall importance of the conclusions of this article is perhaps best commenced by characterizing the audience it's aimed at – a portrait of a sympathetic reader. I take it that such a reader, finding themselves with externalist epistemic intuitions, has a deep sympathy towards Williamson's *MAR*-style analysis of knowledge, and finds *KK* implausible to begin with. The reader may then feel the temptation to spill this skepticism over all introspection principles, thereby rejecting their important practical applications and the intuition that the knowledge of one's inner workings is privileged and authoritative. In the paper, I argued that this temptation can be resisted, and important results and intuitions saved, on principled grounds without steering too far away from Williamson's externalist framework. In section 2, I provided a natural way of restricting Williamson's *MAR* principle (which I took to be a natural explication of externalist intuitions) to first-order knowledge and showed how this restriction allows for both rejecting *KK* and maintaining *KB*, a combination of views labeled *∼KK+KB* thesis. I then proceeded to argue (in sections 3 and 4) that Alex Byrne's transparency framework provides sound philosophical grounds for *∼KK+KB*. *Contra* Das and Salow (2018), I argued that an independently plausible externalist modification of Byrne's theory does a good job of explaining why *KK* might fail, but *KB* remains true. At the end of section 4, I also argued that defending *∼KK+KB* is not only motivated as an expression of certain externalist intuitions but has practical import: it allows to save important applications of introspection principles in dealing with doxastic and epistemic paradoxes and explanation of cooperative or competitive behavior.

Of course, the externalist reader might still object to the proposed restriction of the margin-for-error analysis or challenge the choice of Byrne's framework as arbitrary.

[21]To see that, assume that *a* and *b* both come to know that the other believes *p* and further assume that their epistemic situation is symmetric. Since, by *KB*, they are also in a position to know that they know that they believe that *p* ($K_aK_aB_ap$ *and* $K_bK_bB_bp$), they may, by symmetry, conclude that the other person knows that of them ($K_aK_bB_ap$ *and* $K_bK_aB_bp$). This procedure may be repeated indefinitely, guaranteeing access to any relevant "layer" of epistemic states.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

As explained, however, this combination is quite natural for the epistemic externalist, as it grounds the epistemic success of doxastic self-knowledge in the reliability of transparent inferences without thereby validating *KK* and rejecting the *MAR*-based analysis of knowledge in its plausible applications to first-order knowledge. I hope that by the end of this paper, the externalist reader would be content to agree that, while "[t]o know that you know something is to perform a ( . . . ) great epistemological feat, which is not comparable to just knowing it" (Kripke 2011, p. 34), knowing that you believe requires no such hard work and that at least some of the important practical consequences of introspection principles are saved by defending *KB*.

**Acknowledgments.** I would like to thank Alex Byrne, Tadeusz Ciecierski, and Yonathan Fiat, as well as a reviewer for this journal, for providing extensive comments on previous versions of the manuscript, incorporating which made it significantly better. For additional comments, I want to also thank the audiences in Warsaw and Riga. I want to also thank the Polish-US Fulbright Commission, who facilitated my stay at MIT during the Fall semester of 2023, where I worked on much of the content of the present article.

**Competing interests.** I declare no conflict of interest.

## References

**Antony, L.** (2004). 'A Naturalized Approach to the "A Priori".' *Philosophical Issues* **14**, 1–17.
**Armstrong, D.** (1968). *A Materialist Theory of the Mind*. London: Routledge.
**Audi, R.** (1994). 'Dispositional Beliefs and Dispositions to Believe.' *Noûs* **28**(4), 419–434.
**Bird, A. and Pettigrew, R.** (2021). 'Internalism, Externalism, and the KK Principle.' *Erkenntnis* **86**, 1713–1732.
**Boyle, M.** (2011). 'Transparent Self-Knowledge.' *Aristotelian Society Supplementary Volume* **85**(1), 223–241.
**Burge, T.** (1996). 'Our Entitlement to Self-Knowledge.' *Proceedings of the Aristotelian Society* **96**, 117–158.
**Byrne, A.** (2005). 'Introspection.' *Philosophical Topics* **33**(1), 79–104.
**Byrne, A.** (2012). Knowing What I See. In D. Smithies and D. Stoljar (eds.), *Introspection and Consciousness*, pp. 183–210. Oxford: Oxford University Press.
**Byrne, A.** (2018). *Transparency and Self-Knowledge*. Oxford: Oxford University Press.
**Castañeda, H.N.** (1970). 'On Knowing (or Believing) That One Knows (or Believes).' *Synthese* **21**(2), 187–203.
**Conee, E.** (2005). The comforts of home. *Philosophy and Phenomenological Research* **70**(2), 444–451.
**Das, N. and Salow, B.** (2018). 'Transparency and the KK Principle.' *Noûs* **52**(1), 3–23.
**Dokic, J. and Égré, P.** (2009). 'Margin for Error and the Transparency of Knowledge.' *Synthese* **166**, 1–20.
**Evans, G.** (1982). *Varieties of Reference*. Oxford: Clarendon Press.
**Fagin, R., Halpern, J.Y., Moses, Y. and Vardi, M.** (2004). *Reasoning About Knowledge*. Cambridge, MA: MIT Press.
**Gallois, A.** (1996). *The World Without, the Mind Within: An Essay on First-Person Authority*. Cambridge: Cambridge University Press.
**Gertler, B.** (2012). Renewed Acquaintance. In D. Smithies and D. Stoljar (eds), *Introspection and Consciousness*, pp. 89–123. Oxford: Oxford University Press.
**Gertler, B.** (2021). Self-Knowledge. In E.N. Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). URL = <https://plato.stanford.edu/archives/win2021/entries/self-knowledge/>.
**Greco, D.** (2014). 'Could kk Be Ok?' *The Journal of Philosophy* **111**(4), 169–197.
**Greco, D.** (2015a). 'Iteration Principles in Epistemology I: Arguments for.' *Philosophy Compass* **10**(11), 754–764.
**Greco, D.** (2015b). 'How I Learned to Stop Worrying and Love Probability 1.' *Philosophical Perspectives* **29**, 179–201.
**Harsanyi, J.** (1967) 'Games of Incomplete Information Played by Bayesian Players. Parts I, II, III.' *Management Science* **14**, 159–182, 320–334, 486–502.
**Heal, J.** (1978). Common Knowledge. *The Philosophical Quarterly* **28**(111), 116–131.

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press

**Hintikka, J.** (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca: Cornell University Press.
**James, W.** (1890). *The Principles of Psychology, Volume 1*. New York: Henry Holt and Company.
**Kripke, S.A.** (2011). *Philosophical Troubles: Collected Papers, Volume 1* (Vol. **1**). Oxford: Oxford University Press.
**Lederman, H.** (2018). Uncommon Knowledge. *Mind* **127**(508), 1069–1105.
**McHugh, C.** (2010). Self-Knowledge and the KK Principle. *Synthese* **173**(3), 231–257.
**Moran, R.** (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.
**Nichols, S. and Stich, S.** (2003). *Mindreading: An Integrated Account of Pretence, Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.
**Okasha, S.** (2013). On a Flawed Argument against the KK Principle. *Analysis* **73**(1), 80–86.
**Ramachandran, M.** (2012). The KK-Principle, Margins for Error, and Safety. *Erkenntnis* **76**, 121–136.
**Russell, B.** (1912). *Problems of Philosophy*. New York: Henry Holt and Company.
**Silins, N.** (2012). Judgment as a Guide to Belief. In D. Smithies and D. Stoljar (eds), *Introspection and Consciousness*, pp. 295–328. Oxford: Oxford University Press.
**Smithies, D.** (2019). *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.
**Stalnaker, R.** (1984). *Inquiry*. Cambridge: Cambridge University Press.
**Stalnaker, R.** (2006). On Logics of Knowledge and Belief. *Philosophical Studies* **128**(1), 169–199.
**Tarnowski, M.** (2025). Moore's Paradox and Doxastic Innocence. *Erkenntnis*, Online First, 1–25.
**Weatherson, B.** (2004). Luminous Margins. *Australasian Journal of Philosophy* **82**(3), 373–383.
**Williamson, T.** (1994). *Vagueness*. London: Routledge.
**Williamson, T.** (1996). Cognitive Homelessness. *The Journal of Philosophy* **93**(11), 554–573.
**Williamson, T.** (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.
**Williamson, T.** (2014). Very Improbable Knowing. *Erkenntnis* **79**, 971–999.
**Yalcin, S.** (2024). Defining Common Ground. *Linguistics and Philosophy* **47**(6), 1045–1070.

**Maciej Tarnowski** is an Assistant Professor at the Institute of Philosophy, Jagiellonian University in Kraków. He holds a PhD in philosophy granted by the University of Warsaw. His primary research interests lie at the intersection of epistemology and philosophy of language and the use of formal and experimental methods in these fields.

**Cite this article:** Tarnowski M. (2026). "Knowing That One Believes Without Knowing That One Knows." *Episteme* 1–19. https://doi.org/10.1017/epi.2026.10108

https://doi.org/10.1017/epi.2026.10108 Published online by Cambridge University Press