Knowledge Second*

Adam Bjorndahl

Abstract: Classical philosophical analyses seek to explain knowledge as deriving from more basic notions. The influential "knowledge first" program in epistemology reverses this tradition, taking knowledge as its starting point. From the perspective of epistemic logic, however, this is not so much a reversal as it is the default—the field arguably *begins* with the specialization of "necessity" to "epistemic necessity"—that is, it begins with knowledge.

In this context, putting knowledge *second* would be the reversal. This article motivates, develops, and explores such a "knowledge second" approach in epistemic logic, founded on distinguishing what a body of evidence *actually* entails from what it is (merely) *believed* to entail. We import a logical framework that captures exactly this distinction, use it to define formal notions of (internal and external) justification and knowledge, and investigate applications to the KK principle, the "strong belief" postulate, and the regress problem.

1 Epistemic Logic

Absent a formal underpinning, even good ideas can be difficult to develop and frustrating to communicate. At the other extreme, technical work pursued for its own sake suffers from precisely the same limitations. Formal epistemology, at its best, balances these approaches, each amplifying the other, tapping into a broad and fruitful tradition of bringing mathematical tools to bear on epistemological questions.

The particular species of formalism that concerns us here is the use of modal logic to study *knowledge* and the related concepts of belief, justification, and evidence. This enterprise, often subsumed under the heading "epistemic logic," has its roots in the pioneering work of Hintikka (1962) and has flourished in the decades since through the contributions of countless philosophers and logicians who have extended and enriched this

*This paper was awarded the 2020 Res Philosophica Essay Prize for best unsolicited paper in the special issue on modal epistemology.

Res Philosophica, *Vol. 97, No. 4, October 2020, pp. 471–487*
https://doi.org/10.11612/resphil.1948

paradigm. To set the stage, it will be useful to briefly survey two relevant examples of such contributions.

In "On Logics of Knowledge and Belief," Stalnaker (2006) takes a syntactic approach to investigating the relationship between knowledge and belief. His starting point is a basic propositional language augmented with modalities K and B standing for knowledge and belief, respectively; within this framework, certain assumptions about these attitudes, including relationships between them, can be articulated explicitly as logical axioms.[1] For instance, Stalnaker assumes that knowledge is positively introspective—that is, that knowing φ implies knowing that you know φ, the infamous "KK principle." This is syntactically rendered as $K\varphi \to KK\varphi$. The full list of axioms Stalnaker takes on board (omitting propositional tautologies) is provided for reference in Table 1.

(K_K)	$K(\varphi \to \psi) \to (K\varphi \to K\psi)$	Distribution of knowledge
(T_K)	$K\varphi \to \varphi$	Factivity of knowledge
(4_K)	$K\varphi \to KK\varphi$	Positive introspection for knowledge
(D_B)	$B\varphi \to \neg B\neg\varphi$	Consistency of belief
(sPI)	$B\varphi \to KB\varphi$	Strong positive introspection
(sNI)	$\neg B\varphi \to K\neg B\varphi$	Strong negative introspection
(KB)	$K\varphi \to B\varphi$	Knowledge implies belief
(SB)	$B\varphi \to BK\varphi$	Strong belief

TABLE 1. Stalnaker's axioms

This list includes relatively uncontroversial assumptions, such as (T_K) ("you only know true things") and (KB) ("what you know you also believe"), along with more controversial entries including the aforementioned KK principle (4_K), and (SB), the "strong belief" postulate: "what you believe you believe you know." Regardless of how widely accepted or contested these individual principles may be, the logical framework in which they are embedded allows us to reason about them at a *system* level: What follows from these assumptions, taken together, and what doesn't? Stalnaker proves, for example, that in this system belief is *reducible* to knowledge in the sense that every statement about belief is provably equivalent to a statement just about knowledge.[2]

The second pertinent example is found in "Gettier Cases in Epistemic Logic," where Williamson (2013) takes a semantic approach to studying the relationship between justified true belief and knowledge. His formalization

[1] Or, typically, axiom schemes.

[2] Specifically, Stalnaker proves that $B\varphi \leftrightarrow \neg K\neg K\varphi$ can be derived from the given axioms together with two rules of inference: modus ponens (from φ and $\varphi \to \psi$ infer ψ) and K-necessitation (from φ infer $K\varphi$).

is based on *epistemic models*, which include a set of possible worlds, W, together with a binary relation R on W representing *epistemic accessibility*. Roughly speaking, possible worlds represent the various ways things may be, and a world w' is accessible from a world w, written wRw', just in case w' is compatible with what the agent knows at w. Dually, the agent is said to know φ at w just in case φ is true at all the worlds accessible from w.

Constraints on epistemic models are reflected in corresponding properties of knowledge (or other attitudes). For instance, Williamson adopts the standard assumption that the relation R is reflexive,[3] which corresponds to assuming that the actual world is always compatible with the agent's knowledge at that world, or equivalently, that the agent only knows true things. He also extends the basic models to include a doxastic accessibility relation S to capture belief, and focuses on a special class of models in which worlds encode both "appearance" and "reality." Within this framework Williamson argues for several further constraints, such as the "transparency" of appearance (i.e., the agent is never uncertain about how things appear), and establishes a variety of results, such as the failure of the KK principle and, more centrally, the guaranteed existence of worlds in which the agent has justified true belief without knowledge.

Of course the two approaches just sketched—syntactic and semantic—are quite complementary and frequently presented hand-in-hand. Formulas of epistemic modal languages can be interpreted in appropriately rich epistemic models, and the notion of *provability* (from a set of axioms) is linked to the notion of *validity* (in a class of models) via soundness and completeness theorems. To be sure, Stalnaker does not ignore the semantic interpretations of his axioms, nor does Williamson eschew syntactic expressions of knowledge (or belief); both authors draw freely from the mathematical representations that suit them best in the moment. The difference is one of overall orientation. Stalnaker articulates his assumptions syntactically, as axioms, and explores their syntactic entailments, whereas Williamson argues for adopting certain modeling assumptions and draws out the model-theoretic consequences of those assumptions.

It is what these approaches have in common that is the point of departure for the present work; both take knowledge on board as a kind of primitive—either as a modality, as in Stalnaker's paper, or through the accessibility relation, as in Williamson's. This is certainly a reversal of the classical, philosophical analysis of knowledge as deriving somehow from belief plus something extra, and indeed this reversal has been explicitly recognized and characterized as such by Williamson (2013) under the influential "knowledge first" program in epistemology. From the perspective of epistemic logic, however, this is not so much a reversal as it is the default. The field arguably *begins* with the specialization of the modal concept of

[3]That is, for all $w \in W$, wRw.

"necessity" to "epistemic necessity"; that is, it begins with knowledge (see Hintikka 1962).

Thus, putting knowledge *second* would be the reversal as far as epistemic logic is concerned. This is the track I follow in this article. I begin by motivating a logical framework that includes neither knowledge nor epistemic accessibility as a primitive; nonetheless, within this framework something we might reasonably call knowledge can be derived by combining other primitive concepts. This allows for a formal analysis of the properties of knowledge that bottoms out not in abstract modelling assumptions or axiomatics, but instead in features of the constitutive concepts of belief and evidence.

2 Internalism, Externalism, and Evidence

The classic philosophical debate between epistemic internalism and externalism, and in particular attendant questions regarding the nature of justification and evidence, provides a convenient entry-point into the "knowledge-second" logical models that I wish to motivate and develop.

If I say "Anne has a justified belief that p," one thing I might mean is that Anne ought not be held in poor regard if p turns out to be false. In this case, what is justified is Anne's *having* the belief that p, and this justification speaks positively of her—I am saying something good about Anne's relationship to her belief-that-p. Of course, such a gloss is silent on many details, and how one spells out these details can give rise to a multitude of quite different theories. Nonetheless, presenting justification as a property of the relationship between an agent and their beliefs has a distinctly internalist flavor to it; any *actual* connection between Anne's belief-that-p and the external fact-of-p is not (directly) relevant to assessing whether she is justified. Demons do not disturb us here.

By contrast, an externalist reading of "Anne has a justified belief that p" might parse "justified" as applying more narrowly (syntactically speaking): what is justified is the *belief* that p, itself—the one that Anne happens to have. This can naturally be understood as saying something good about the relationship between the belief-that-p and the fact-of-p. Anne may well have done something praiseworthy to come to hold such a belief, but regardless, the claim of justification here rests ultimately not on how Anne has comported herself, but on whether the belief she has actually bears an appropriate relationship to the world.

Once again, many details are omitted on such a cursory account, yielding a corresponding abundance of broadly externalist elaborations. Moreover, I make no claim that these two tales of what justification might mean exhaustively partition the relevant conceptual space. However, they do provide the scaffolding for a distinction that we can formalize, and through that formalization, draw substantial insight.

On both perspectives, justification is taken to consist in a relationship between the belief-that-p and something else. Our externalist demands a relationship to the fact-of-p (i.e., to the "external world"), while our internalist demands a relationship to the agent (i.e., something "internal" to Anne). But this level of description is far too coarse to admit much analysis. To make progress, we need to incorporate some additional structural assumptions, the goal being to narrow our focus to the point where the issues become tractable while remaining, ideally, as general as possible. To this end, we turn to the concept of *evidence*, and specifically the role of evidence in supporting justification.

It is commonplace in both casual discourse and in careful philosophical presentations to interpret justification, implicitly or explicitly, as deriving from evidence. In Gettier's (1963) seminal critique of knowledge as justified true belief, for example, we are provided multiple scenarios where Smith has "strong evidence" for some proposition or another. Smith's evidence for Jones having ten coins in his pocket is Smith's having recently counted them; his evidence for Jones being the man who will get a certain job is an assurance from the president of the company, etc. Similarly, in Goodman's (1976) famous "fake barn" scenario, "evidence" is treated as essentially synonymous with "justification"—the evidence for the barn is Henry's perceptual experience of seeing the barn, and this in turn justifies Henry's belief that there is a barn before him. For a more recent example, take the *Oxford Handbooks Online* entry for "Formal Epistemology," where we find the following introductory description of the internalist position:

> What does it take for me to be epistemically justified in believing something? The most compelling and common answers assert that you have a justified belief when this belief rests upon your internally *having sufficient evidence or reason for that belief* (Douven and Schupbach 2017, emphasis mine).

So let's take for granted that there is such a thing as "evidence," that agents can obtain evidence in certain ways, that evidence can entail propositions about the world, and that justification has something to do with there being the "right" sort of evidence. How can we unpack the internalist and externalist views of justification we have been considering, in terms of evidence?

The externalist position admits a fairly immediate, intuitive analysis: the required relationship to the "fact-of-p" is simply that Anne's evidence, call it e, *actually* entails p. That is, Anne's belief is (externally) justified (henceforth: ex-justified) exactly when it is a belief in something that is, in fact, entailed by her evidence. I do not claim that this is the only available analysis of ex-justification in terms of evidence entailment, only that it is a reasonably natural one. Of course, it remains silent on many details,

but as we will see it is already contentful enough to support a fruitful investigation.

Where does this leave the internalist? Whether e actually entails p is, presumably, a fact about the world, not something internal to Anne. All we are assuming about Anne's relationship to e is that she "has" it, in some sense—in the sense that Smith who has counted the coins in Jones' pocket "has" that evidence, or that Henry who is receiving barn-like visual impressions "has" that evidence. But barn-impressions do not necessarily entail the existence of a barn; this is the whole point of the fake barn scenario. On the other hand, Henry may certainly be under the impression that his barn-impressions entail the existence of a barn; similarly, Anne may *believe* that her evidence e entails p. And since this belief of Anne's—possibly false, possibly true-but-misguided, misleading, etc.—is internal to her, it provides a natural candidate for a notion of internalist justification; Anne's belief is (internally) justified (henceforth: in-justified) exactly when it is a belief in something that she believes is entailed by her evidence.

On this view, Henry, driving along in fake-barn country, probably is in-justified in his belief that there is a barn in front of him, because (presumably) he genuinely believes that his sense-impressions entail the existence of the barn. Whether he is ex-justified in his belief about the barn is more subtle; it depends on whether his sense-impressions in that scenario *actually* entail the existence of the barn, which in turn depends on your theory of what constitutes "actual evidence entailment." Crucially, we need not specify such a theory to continue our investigation; we already have enough raw material to develop a general logical model for belief, evidence, and justification with interesting applications to the internalism/externalism debate and our understanding of knowledge more generally.

3 Models for Evidence

Like Williamson, we take a primarily semantic approach to developing a logic appropriate for reasoning about belief and evidence, although we help ourselves to syntactic formulations wherever it aids the exposition. To build appropriate models, we need to set up a mathematical framework rich enough to capture the conceptual primitives we have laid out, which fall into two categories: (1) the beliefs of some fixed agent and (2) evidence and evidence-entailment.

We begin with the latter. Building a logical model for evidence is not a novel enterprise; many such models exist. Most of them identify a piece of evidence with a set of possible worlds, $U \subseteq W$, and interpret evidence entailment via set containment: U entails exactly those propositions $p \subseteq W$ such that $U \subseteq p$ (see, for example, Parikh et al. 2007; Benthem and Pacuit 2011; Baltag et al. 2016).

This model simply will not do for our purposes. Indeed, the inadequacy of this mathematical treatment of evidence for reasoning about the

internalist and externalist positions described earlier is at the core of this paper. A key distinction we seek to capture, that between in-justification and ex-justification, is predicated on the distinction between a piece of evidence actually entailing some proposition versus an agent (merely) believing that it does. For this contrast to be meaningful, it must be possible for an agent (at least in principle) to be mistaken about what they believe a given piece of evidence entails; however, in a model where evidence is identified with sets of possible worlds and entailment with containment, this is impossible. Whether or not $U \subseteq p$ is not something the agent can be uncertain about, it is simply a fact about the model.

To overcome this limitation, we need to generalize the representation of evidence, and for this we adopt a framework recently proposed by Bjorndahl and Özgün (2019) in which, rather than being identified with a single subset of W, a body of evidence can have multiple "interpretations" as different subsets of W, depending on the world. More precisely, in addition to a set of possible worlds W, an evidence model includes a set of evidence states $\mathcal{E}$ together with interpretations $I_e : W \to 2^W$,[4] one for each $e \in \mathcal{E}$. Intuitively, elements of $\mathcal{E}$ represent "total states of evidence" that the agent may find themselves in, while $I_e(w)$ represents the "actual" or "correct" interpretation of e at world w. Evidence entailment is still given by set containment, but it is now evaluated relative to a world w: given a proposition $p \subseteq W$, we say that e actually entails p at w just in case $I_e(w) \subseteq p$.

Note that, as promised, we do not need to have a theory of "actual evidence entailment" in hand to work with these models. Effectively, actual evidence entailment is treated as a primitive, in much the same way that Stalnaker and Williamson treat knowledge. We can put constraints on it—that is, we can demand that it satisfy certain conditions (and we will do so)—but we need not have an explanation of where the interpretations I_e "come from." They are simply stipulated. Far from a cheat, this is one of the fundamental benefits of the mathematical/logical approach to philosophical questions; it allows for an investigation of the properties of a given concept (in this case, evidence entailment), and its interplay with other concepts, while remaining agnostic about its internal structure. Of course, such systematic agnosticism is available in any context, not just formal analyses; however, formalism arguably emphasizes, even forces, this tactic. When one is obliged to be completely precise about *every* component of a theory, one must explicitly identify some of those components as primitives.

We summarize the key statement "e actually entails p at w" with a single bit of notation: $(w, e) \models E p$. Effectively, this amounts to introducing a new modality E for "actual evidence entailment," and interpreting this modality not with respect to worlds, but world-evidence pairs. Following

[4] We write 2^X to denote the *powerset* of X—that is, the set of all subsets of X.

Bjorndahl and Özgün (2019), we restrict our attention to those world-evidence pairs that are *coherent* in the sense that $w \in I_e(w)$, and call such pairs *evidence scenarios*. In other words, we eliminate those world-evidence pairs (w,e) where the actual interpretation of e at w rules out w. This captures a kind of factivity assumption about evidence: all evidence scenarios satisfy $(w,e) \models E\,p \to p$, which says that whatever the evidence actually entails is true. It will also be convenient to collect together, for each $e \in \mathcal{E}$, the set of worlds coherent with e in the sense defined above; thus, we define $U_e = \{w \in W : w \in I_e(w)\}$.

Evidence models do not come equipped with epistemic accessibility relations, since knowledge is not meant to be a primitive of this system. However, we do wish to encode belief; therefore, we assume that the agent's beliefs at (w,e) are given by specifying a nonempty subset $V \subseteq U_e$, thought of as the *doxastically accessible worlds*. Intuitively, these are the worlds that are compatible with the agent's beliefs. The assumption that $V \neq \emptyset$ corresponds to the *consistency* condition for beliefs (i.e., axiom (D_B) on Table 1), while the assumption that $V \subseteq U_e$ ensures that the agent does not consider incoherent evidence scenarios possible.

The preceding discussion of evidence models and their semantics is formalized as follows. An **evidence model** is a tuple $(W, \mathcal{E}, I)$ where W and $\mathcal{E}$ are nonempty sets and $I = \{I_e\}_{e \in \mathcal{E}}$ is a parametrized family of functions $I_e : W \to 2^W$. Our language is a bimodal extension of the basic propositional language obtained by including the modalities E and B; primitive propositions (p, q, r, etc.) are identified with subsets of W in the usual way. Truth of formulas in this language is evaluated with respect to triples $(w,e,V) \in W \times \mathcal{E} \times 2^W$, where $w \in I_e(w)$ and $\emptyset \neq V \subseteq U_e$, according to the following recursive clauses:

$$
\begin{array}{lcl}
(w,e,V) \models p & \Leftrightarrow & w \in p \\
(w,e,V) \models \neg\varphi & \Leftrightarrow & (w,e,V) \not\models \varphi \\
(w,e,V) \models \varphi \wedge \psi & \Leftrightarrow & (w,e,V) \models \varphi \text{ and } (w,e,V) \models \psi \\
(w,e,V) \models E\varphi & \Leftrightarrow & I_e(w) \subseteq [\![\varphi]\!]^{e,V} \\
(w,e,V) \models B\varphi & \Leftrightarrow & V \subseteq [\![\varphi]\!]^{e,V},
\end{array}
$$

where $[\![\varphi]\!]^{e,V} := \{w \in W : (w,e,V) \models \varphi\}$.

The notation $[\![\cdot]\!]^{e,V}$ may be unfamiliar; it simply serves to transform each formula φ into a proposition (i.e., a subset of W: $[\![\varphi]\!]^{e,V} \subseteq W$). The superscript is necessary to keep track of the fact that the interpretation of φ may depend on the evidence state or the doxastically accessible worlds.

Evidence models validate a number of simple properties of belief and evidence entailment, some of which we have already noted, such as consistency of belief (D_B) and factivity of evidence entailment (T_E). For convenience, we collect these core properties in Table 2.

(K_E)	$E(\varphi \to \psi) \to (E\varphi \to E\psi)$	Distribution of evidence entailment
(T_E)	$E\varphi \to \varphi$	Factivity of evidence entailment
(K_B)	$B(\varphi \to \psi) \to (B\varphi \to B\psi)$	Distribution of belief
(D_B)	$B\varphi \to \neg B\neg\varphi$	Consistency of belief
(4_B)	$B\varphi \to BB\varphi$	Positive introspection for belief
(5_B)	$\neg B\varphi \to B\neg B\varphi$	Negative introspection for belief
(EB)	$B\varphi \to EB\varphi$	Evidence for belief
(ENB)	$\neg B\varphi \to E\neg B\varphi$	Evidence for nonbelief

TABLE 2. Basic properties of belief and evidence entailment

It is worth emphasizing that the recursive clauses for E and B encode certain "transparencies" involving the agent's evidence and her beliefs. For one thing, belief is treated as fully introspective, satisfying both (4_B) and (5_B). As with many assumptions hardcoded in such logical models, there are significant idealizations involved. While we can always work to generalize our models away from specific idealizations, in this particular case assuming that belief is fully introspective serves a useful purpose. We will subsequently be investigating various notions of *justified belief*, and a pertinent question will be whether these notions satisfy introspection conditions. Of course, one easy way for justified belief to fail to be introspective is on account of belief itself lacking this property. So, naturally, we are interested in the question of whether introspection properties are *preserved* when belief is upgraded to justified belief, which is a question we can address in the present framework.

The other type of transparency given in Table 2 is captured by (EB) and (ENB), which together tell us that the agent always has evidence that they have the beliefs they actually have. Note that this is certainly not to say that the agent always has evidence that *supports* their beliefs, which would instead be written $B\varphi \to E\varphi$. Rather, we are effectively assuming that the agent's total evidence includes their own introspective access to their beliefs, so that whenever they (don't) believe φ, they necessarily have evidence that entails that they (don't) believe φ. Stalnaker's principles of strong positive and negative introspection (see Table 1)—(sPI) and (sNI)—can be seen as analogues to (EB) and (ENB), respectively, with knowledge K in place of evidence entailment E.

4 Defining Knowledge

We are now in a position to formalize the evidential conceptions of internal and external justification articulated previously. Recall that ex-justification for a belief in p simply requires that, in addition to the agent believing p,

p actually follows from their evidence. So we can say that the agent has an ex-justified belief in p at (w,e,V) just in case $V \subseteq p$ and $I_e(w) \subseteq p$. To denote this we write $(x,e,V) \models B^{ex} p$, and more generally we define

$$(w,e,V) \models B^{ex}\varphi \quad \Leftrightarrow \quad V \subseteq [\![\varphi]\!]^{e,V} \text{ and } I_e(w) \subseteq [\![\varphi]\!]^{e,V}.$$

Internal justification is a bit trickier; we want to say that the agent is in-justified in believing p just in case they believe p and they believe that their evidence entails p. Thus, the agent is in-justified in believing p at (w,e,V) just in case $V \subseteq p \cap [\![Ep]\!]^{e,V}$. To denote this, we write $(w,e,V) \models B^{in} p$. In fact, factivity of E guarantees $[\![Ep]\!]^{e,V} \subseteq p$, and indeed $[\![E\varphi]\!]^{e,V} \subseteq [\![\varphi]\!]^{e,V}$, so in general we define

$$(w,e,V) \models B^{in}\varphi \quad \Leftrightarrow \quad V \subseteq [\![E\varphi]\!]^{e,V}.$$

It is not difficult to see that under these definitions we can derive the equivalences presented in Table 3. These expressions align nicely with some of the high-level, pre-theoretic intuitions about external and internal justification highlighted previously. An ex-justified belief $B^{ex}\varphi$ demands an objective component—namely $E\varphi$: the agent's evidence must actually stand in an appropriate (i.e., entailing) relationship to the fact-of-φ. By contrast, an in-justified belief $B^{in}\varphi$ is totally subjective; it consists entirely in the agent having a certain kind of belief. While the required belief is stronger than simply believing φ (one easily checks that $BE\varphi \to B\varphi$, using factivity of E), nonetheless it is merely a belief and nothing more.

(Ext)	$B^{ex}\varphi \leftrightarrow (B\varphi \wedge E\varphi)$	External justification
(Int)	$B^{in}\varphi \leftrightarrow BE\varphi$	Internal justification

TABLE 3. External and internal justification

Ex-justified belief is factive ($B^{ex}\varphi \to \varphi$) but *not* positively introspective ($B^{ex}\varphi \not\to B^{ex}B^{ex}\varphi$); as such, on a first pass it coincides well with the externalist conception of knowledge, which is typically taken to violate the KK principle. Indeed, the rejection of KK is often presented as practically immediate from the externalist position. For instance, as Dretske (2004) explains, externalist knowledge "depends for its existence on circumstances of which the knower may be entirely ignorant. So the knower can know that P without knowing (as required by KK) that he knows that P" (176).

Let us then provisionally adopt B^{ex} as our (externalist) candidate for knowledge and, in this context, unpack the failure of positive introspection and compare it to the general externalist rationale for rejecting KK. Note that this "unpacking" is available to us precisely because we did not take

knowledge (or ex-justification) as a primitive in our models, so whatever properties knowledge turns out to have (or lack) necessarily derive from more fundamental features of the system.

From (Ext) and some basic modal reasoning we can see that

$$\begin{aligned} B^{ex}B^{ex}\varphi &\leftrightarrow B^{ex}(B\varphi \wedge E\varphi) \\ &\leftrightarrow B(B\varphi \wedge E\varphi) \wedge E(B\varphi \wedge E\varphi) \\ &\leftrightarrow BB\varphi \wedge BE\varphi \wedge EB\varphi \wedge EE\varphi. \end{aligned}$$

Thus, the KK principle in this setting can be expressed as follows:

$$(B\varphi \wedge E\varphi) \rightarrow (BB\varphi \wedge BE\varphi \wedge EB\varphi \wedge EE\varphi).$$

We have already seen that both $BB\varphi$ and $EB\varphi$ follow from $B\varphi$, so the failure of KK in our models stems from the following two non-implications

(1) $$(B\varphi \wedge E\varphi) \not\rightarrow BE\varphi$$

and

(2) $$(B\varphi \wedge E\varphi) \not\rightarrow EE\varphi,$$

both of which are easily seen to be witnessed.[5]

Thus we are presented with two logically distinct obstacles to KK. When the consequent in (1) fails, we might reasonably describe the problem as residing with the agent—simply knowing φ does not entail that one also believes that their evidence entails φ. This is perhaps a natural way of interpreting Dretske's reference to "circumstances of which the knower may be entirely ignorant" (176) and other externalist objections to KK that place the blame in some form or another on the agent failing to have the right kind of (second-order) beliefs. Specifically, it is entirely possible to believe something that is, as a matter of fact, entailed by the evidence, without also believing that it is entailed by the evidence. Interestingly, since the consequent in (1) is equivalent to our notion of in-justified belief, we can also describe this obstacle to KK as consisting in the fact that ex-justified beliefs may not be in-justified.

On the other hand, when the consequent in (2) fails, we are faced not with a deficiency of the agent, but more a raw fact about the world—the fact that evidence may not be "self-affirming" in the sense of entailing that it entails whatever it in fact entails. This statement is a mouthful, but it arguably corresponds to a commonsense property of measurement. Suppose, for example, that I stand on a scale and read "78 kg" on the

[5]That is, it is easy to construct models in which, for some (w, e, V), we have $(w, e, V) \models B\varphi \wedge E\varphi$ but $(w, e, V) \not\models BE\varphi$ or $(w, e, V) \not\models EE\varphi$.

display; suppose further that the scale is calibrated quite well, and its sensitivity is such that this reading is evidence (i.e., entails) that I in fact weight less than 80 kg. We need not conclude, from these assumptions, that the reading on the scale also entails that it is calibrated in this nice way or has a certain margin of error. Indeed, to measure the sensitivity or calibration of a tool, we typically use some other tool—that is, we seek out further evidence. This suggests that the failure in (2) is quite natural. Moreover, it gives us a rather different vision of the failure of KK as deriving from a more fundamental failure: the failure of evidence to be self-affirming. This view, and the arguments that support it, align closely with the "margin of error" arguments against KK advanced by Williamson (2000).

5 Strong Belief and Full Justification

Turning back to in-justified belief, we can immediately see that it is not a very good candidate for knowledge, since it is not even factive—beliefs can always be mistaken. However, an agent with an in-justified belief of φ does not *feel* like they are mistaken; in fact, they believe that they have *knowledge* of φ (in the sense of ex-justification):

$$B^{in}\varphi \to BB^{ex}\varphi. \tag{3}$$

To see this, first note that $BB^{ex}\varphi$ is equivalent to $BB\varphi \wedge BE\varphi$; (3) then follows easily from (T_E) and (4_B). (The converse also holds, as is easily checked.)

The implication in (3) bears a striking resemblance to Stalnaker's "strong belief" axiom, (SB), where B^{ex} plays the role of K and, instead of a single belief modality, there are two involved: plain belief B and in-justified belief B^{in}. In fact, this is essential: both $B\varphi \to BB^{ex}\varphi$ and $B^{in}\varphi \to B^{in}B^{ex}\varphi$ are strengthenings of (3), and neither is valid in evidence models.[6] This suggests a new perspective on the strong belief postulate, one that could not be articulated in a less expressive logic. After all, the motivating idea of (SB) was to capture a strong notion of belief as "subjective certainty" (Stalnaker 2006, 179); in this spirit, it seems reasonable both to interpret the belief that's supposed to be strong as being something *more* than mere belief (i.e., as in-justified belief, B^{in}), while simultaneously interpreting the "subjective" part of "subjective certainty" less stringently (i.e., as referring to mere belief, B). Thus, an agent has a strong belief in φ just in case they (merely) believe that they know φ.

It is not hard to check that B^{in} is positively introspective. If one believes that the evidence entails φ, then one believes that the evidence

[6] Briefly, the former fails because in general $B\varphi$ does not entail $BE\varphi$, the latter because $BE\varphi$ does not entail $BEE\varphi$.

entails that they believe that the evidence entails φ:

$$BE\varphi \rightarrow BEBE\varphi.$$

This implication is not quite so impressive as a quick reading may make it seem: note that the second "layer" of evidence in the consequent is evidence that entails $BE\varphi$, which effectively comes for free from the "evidence for belief" property of our models, (EB). In other words, as previously discussed, agents are assumed by default to have evidence for their having whatever beliefs they happen to have; positive introspection for in-justified belief does not amount to much more than this.

However, this observation hints at another property that it is illuminating to consider in more detail. Recall that the definition of in-justified belief was supposed to capture a sort of blamelessness on the part of the agent. An in-justified belief in φ is not supposed to be arbitrary; rather, it is a belief that the agent believes is actually supported by her evidence. It is this sense in which the agent has done her due diligence, so to speak; she has tried her best to believe something that follows from her evidence.

But has she really tried her best? Whence comes this belief that her original belief is supported by the evidence? What if this belief was itself formed arbitrarily? Should we not demand that it, too, be supported by the evidence, or at least that the agent believe it to be so supported? Following this line of reasoning, it is tempting to seek out a new, stronger notion of internally justified belief, call it B^*, satisfying the following:

$$B^*\varphi \leftrightarrow B^*E\varphi.$$

Taken as a definition, this would clearly be circular, but that doesn't preclude it from being useful. In a sense, it can be viewed as the "limit" of a procedure that begins with the original definition of B^{in} but then demands that every belief used in the definition of in-justification be in-justified as well. We might picture this as a chain of equivalences ($\leftrightarrow$) and rewrite rules ($\curlyvee$):

$$\begin{array}{ccccccccc}
B^{in}\varphi & \leftrightarrow & BE\varphi & & & & & & \\
 & & \curlyvee & & & & & & \\
 & & B^{in}E\varphi & \leftrightarrow & BEE\varphi & & & & \\
 & & & & \curlyvee & & & & \\
 & & & & B^{in}EE\varphi & \leftrightarrow & BEEE\varphi & & \\
 & & & & & & \curlyvee & & \\
 & & & & & & B^{in}EEE\varphi & \leftrightarrow & \cdots
\end{array}$$

This is not a formal definition, just a picture to build intuition. What it makes clear is that demanding in-justified beliefs to support in-justified beliefs creates a chain of corresponding evidential demands: evidence to support φ (i.e., $E\varphi$), evidence to support that this evidence supports φ (i.e., $EE\varphi$), evidence to support that *this* evidence supports that the original evidence supports φ (i.e., $EEE\varphi$), and so on. So this is a version of the

familiar *regress problem*, dressed up in a new formalism, but the attire here is not merely for show. We have at our disposal, in the evidence model framework, the means to "solve" the regress by finding its "limit" (or, perhaps more perspicuously, its fixed point).

Consider the formula $BE\varphi$: by definition, this holds at (w, e, V) exactly when $V \subseteq [\![E\varphi]\!]^{e,V}$, or equivalently, when every $w' \in V$ satisfies $I_e(w') \subseteq [\![\varphi]\!]^{e,V}$. This in turn is equivalent to demanding that

$$\bigcup_{w' \in V} I_e(w') \subseteq [\![\varphi]\!]^{e,V}.$$

Call this union V_1; thus, V_1 represents the agent's in-justified beliefs in the same way that V represents her beliefs. Moreover, it is easy to see that $V_1 \supseteq V$, which naturally corresponds to the fact that the agent has fewer in-justified beliefs than mere beliefs.

This process can be iterated. If we consider the formula $BEE\varphi$, we see that this holds at (w, e, V) exactly when $V_1 \subseteq [\![E\varphi]\!]^{e,V}$, or equivalently, when every $w' \in V_1$ satisfies $I_e(w') \subseteq [\![\varphi]\!]^{e,V}$. Therefore, if we define

$$V_2 = \bigcup_{w' \in V_1} I_e(w'),$$

then we have $(w, e, V) \models BEE\varphi$ if and only if $V_2 \subseteq [\![\varphi]\!]^{e,V}$. So V_2 represents the second iteration of in-justification.

Recursively define

$$V_n = \bigcup_{w' \in V_{n-1}} I_e(w').$$

This produces a nested increasing chain $V \subseteq V_1 \subseteq \cdots \subseteq V_n \subseteq \cdots$ of sets representing more and more stringent versions of the agent's beliefs, starting from their initial, most liberal beliefs given by V. Let $V^* = \bigcup_n V_n$, and define

$$(w, e, V) \models B^*\varphi \quad \Leftrightarrow \quad V^* \subseteq [\![\varphi]\!]^{e,V}.$$

By construction, we have $B^*\varphi \to BE^k\varphi$ for all $k \geq 0$. I will call $B^*\varphi$ a *fully justified belief* (understanding justification in this context as a species of *internal* justification). One can check that fully justified beliefs are positively introspective ($B^*\varphi \to B^*B^*\varphi$) and "strong" in Stalnaker's sense, if we interpret knowledge as ex-justified belief ($B^*\varphi \to B^*B^{ex}$). Furthermore, it is not hard to show that for each $w' \in V^*$, we have $I_e(w') \subseteq V^*$,[7] making V^* a fixed point of the recursive operation used above to construct the sets V_n. More to the point, this means that the equivalence $B^*\varphi \leftrightarrow B^*E\varphi$ is

[7] If $w' \in V^*$ then $w' \in V_n$ for some n, so $I_e(w') \subseteq V_{n+1} \subseteq V^*$.

valid in evidence models—exactly the "circular" idea we started with, now semantically grounded.

6 Refining Knowledge

Although fully justified belief has several nice properties, it is easy to see that it is not factive, making it a bad candidate for knowledge. On the other hand, despite our provisional adoption, ex-justified belief is also, arguably, a bad candidate for knowledge. I will close by considering some alternatives that naturally present themselves in the present framework.

The conjunction $B\varphi \wedge E\varphi$ certainly seems to align with a purely "external" conception of justification; however, when it comes to upgrading belief to knowledge, one might protest that adding $E\varphi$ is not enough. After all, it could be by pure chance that something you believe to be true also happens to be entailed by your evidence. This requires no self-awareness on the part of the supposed knower—just simple coincidence. Is that really enough for knowledge?

This style of objection has clear internalist overtones, but in the context of the logical system we have developed we can take it seriously without abandoning the externalist point of view altogether. Perhaps the most natural way to do so is also the most straightforward: declare that knowledge requires both external *and* internal justification. Let us take this as the definition of a new modality, K:

$$(w, e, V) \models K\varphi \quad \Leftrightarrow \quad (w, e, V) \models B^{ex}\varphi \wedge B^{in}\varphi.$$

It is not hard to see that this definition yields the following equivalence:

$$K\varphi \leftrightarrow (BE\varphi \wedge E\varphi). \tag{4}$$

Thus, on this account, knowledge of φ amounts to a *true* belief that one's evidence entails φ. In a sense, this has quite a classical flavor to it—we strengthen the clearly inadequate "knowledge = true belief" account by incorporating justification, as usual, but in a different order. Instead of "knowledge = justified, true belief," we have arrived at "knowledge = true belief of justification."

As before, we can investigate the properties of this modality, perhaps with some optimism that its hybrid internalist/externalist nature might result in some interesting departures from the purely externalist candidate we considered previously. For example, in our discussion of the failure of positive introspection for B^{ex}, we noted that one of the two obstacles was precisely the fact that ex-justified beliefs may not be in-justified. This is very much in line with our motivation for defining the K modality. Is this obstacle then circumvented by K?

Perhaps surprisingly, it is not. Using (4), we can check that positive introspection for K is equivalent to

$$(BE\varphi \wedge E\varphi) \to (BEBE\varphi \wedge BEE\varphi \wedge EBE\varphi \wedge EE\varphi).$$

We have already observed that $EBE\varphi$ and $BEBE\varphi$ follow from $BE\varphi$, and that $EE\varphi$ does not follow from $E\varphi$ (this was the second obstacle to positive introspection—the failure of evidence to be "self-affirming"—which, intuitively, we would not expect our definition of K to avoid). But we have also seen that $BEE\varphi$ does not, in general, follow from $BE\varphi$, and it is easy to check that

$$(BE\varphi \wedge E\varphi) \to BEE\varphi$$

is *not* valid in evidence models.

The fact that $BE\varphi \not\to BEE\varphi$ was the core motivation for our development of the "fully justified belief" modality B^*; this suggests that it may be fruitful to consider a correspondingly strengthened notion of knowledge, namely,

$$(w, e, V) \models K^*\varphi \quad \Leftrightarrow \quad (w, e, V) \models B^{ex}\varphi \wedge B^*\varphi,$$

which yields the following equivalences:

$$K^*\varphi \leftrightarrow (B^*\varphi \wedge E\varphi) \leftrightarrow (B^*E\varphi \wedge E\varphi). \tag{5}$$

It is not hard to see now that the only obstacle to positive introspection for K^* is the fact that $E\varphi \not\to EE\varphi$—namely, the failure of evidence to be self-affirming. In other words, although the KK principle still fails when applied to the conception of knowledge captured by the K^* modality (i.e., the implication $K^*\varphi \to K^*K^*\varphi$ is not valid), its failure has nothing to do with any "ignorance" on the part of the knower. Indeed the knower, by assumption, has full confidence that their evidence is sufficient to establish what they know. In fact, evidence models validate $B^*\varphi \to B^*K^*\varphi$—that is, Stalnaker's strong belief postulate is satisfied when B^* and K^* play the role of belief and knowledge, respectively.

From (5) we read that $K^*\varphi$ corresponds to a fully justified, true belief that the evidence entails φ. Or, if we are being careful to keep track of the internal and external senses of justification at play here, we get "knowledge = fully in-justified, true belief in ex-justification." Granted, this is a fairly convoluted series of adjectives, but that is part of the point of this article; the logical formalism we rely on allows us to make subtle distinctions and parse sophisticated expressions like these with ease, extending our ability to reason about core epistemological issues.

7 Reflecting on the Evidence

I have endeavored to show, in a novel way, the value of formalism in approaching questions in epistemology, and in particular the insight that can be gained by "putting knowledge second" in the context of epistemic logic. There is no shortage of existing logics for reasoning about knowledge, belief, justification, and so on. This means there is—or there ought to be—a rather high bar for new additions to the field. By putting knowledge second, and focusing on the role of evidence and subjective uncertainty about what evidence entails, I believe that bar is cleared. The evidence models presented herein offer new insights to core issues in epistemology, both classical and contemporary: from the internalism/externalism debate to the regress problem, from the failure of KK to the very definition of knowledge.

Adam Bjorndahl
Carnegie Mellon University
E-mail: abjorn@cmu.edu

References:

Baltag, Alexandru, Nick Bezhanishvili, Aybüke Özgün, and Sonja Smets. 2016. "Justified Belief and the Topology of Evidence." In *Logic, Language, Information, Computation*, edited by Jouko Väänänen, Åsa Hirvonen, and Ruy Queiroz, Vol. 9803 of *Lecture Notes in Computer Science*. Heidelberg: Springer.

Benthem, Johan and Eric Pacuit. 2011. "Dynamic Logics of Evidence-Based Beliefs." *Studia Logica* 99 (1): 61–92. https://doi.org/10.1007/s11225-011-9347-x.

Bjorndahl, Adam and Aybüke Özgün. 2019. "Uncertainty about Evidence." In *Proceedings of the 17th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, edited by Lawrence S. Moss, 68–81. https://arxiv.org/abs/1907.08335.

Douven, Igor and Jonah N. Schupbach. 2017. *Formal Epistemology*. Oxford: Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/.

Dretske, Fred. 2004. "Externalism and Modest Contextualism." *Erkenntnis* 61 (2/3): 173–186. https://doi.org/10.1007/s10670-004-9277-3.

Gettier, Edmund. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23 (6): 121–123. https://doi.org/10.1093/analys/23.6.121.

Goodman, Alvin I. 1976. "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73 (20): 771–791. https://doi.org/10.2307/2025679.

Hintikka, Jaakko. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Parikh, Rohit, Lawrence S. Moss, and Chris Steinsvold. 2007. "Topology and Epistemic Logic." In *Handbook of Spatial Logics*, edited by Marco Aiello, Ian Pratt-Hartmann, and Johan van Benthem, 299–341. Netherlands: Springer.

Stalnaker, Robert. 2006. "On Logics of Knowledge and Belief." *Philosophical Studies* 128 (1): 169–199. https://doi.org/10.1007/s11098-005-4062-y.

Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Williamson, Timothy. 2013. "Gettier Cases in Epistemic Logic." *Inquiry* 56 (1): 1–14. https://doi.org/10.1080/0020174X.2013.775010.