The Logic of Justified Belief, Explicit Knowledge, and Conclusive Evidence

Alexandru Baltag Bryan Renne* Sonja Smets†

University of Amsterdam
Institute for Logic, Language, Information and Computation (ILLC)

Abstract

We present a complete, decidable logic for reasoning about a notion of *completely trustworthy ("conclusive") evidence* and its relations to *justifiable* (implicit) belief and knowledge, as well as to their *explicit justifications*. This logic makes use of a number of evidence-related notions such as availability, admissibility, and "goodness" of a piece of evidence, and is based on an innovative modification of the Fitting semantics for Artemov's Justification Logic designed to preempt Gettier-type counterexamples. We combine this with ideas from belief revision and awareness logics to provide an account for explicitly justified (defeasible) knowledge based on conclusive evidence that addresses the problem of (logical) omniscience.

1 Introduction

Justification Logic, due to Sergei Artemov and originally conceived as a solution to a long-standing open problem concerning the intended semantics of Gödel's provability logic [3], has since developed into a wide-ranging study of the notions of *evidence* and *justification*; see, e.g., [1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 18, 19, 20, 21, 23, 24, 25, 28, 30, 31, 33, 42, 43, 44, 45, 50, 53, 54, 59, 60, 61, 62]. By making explicit the "evidence" supporting a given assertion, this formalism can capture one of the main ingredients in the epistemological analysis of knowledge: the *epistemic justification* underlying the knowledge or belief possessed by an agent endowed with only limited logical resources and bounded rationality. As a consequence, this approach can be used to address the problem of "logical omniscience" [11, 12, 28, 31, 59, 60] that affects other formalizations of doxastic/epistemic logic; indeed, Justification Logic is perhaps the most philosophically far-reaching and computationally sophisticated approach to this well-known problem.

NOTICE: this is the authors' version of a work that was accepted for publication in *Annals of Pure and Applied Logic*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Annals of Pure and Applied Logic*, 165(1):49–81, 2014. doi:10.1016/j.apal.2013.07.005

*Funded by an Innovational Research Incentives Scheme Veni grant from the Netherlands Organisation for Scientific Research (NWO).

†Funded in part by an Innovational Research Incentives Scheme Vidi grant from the Netherlands Organisation for Scientific Research (NWO) and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 283963.

Melvin Fitting's semantics [10, 29] for this logic relates in an elegant way to Kripke's well-known relational semantics for knowledge and belief: essentially, in the Fitting semantics, a piece of evidence t justifies the belief in (or knowledge of) an assertion φ iff t is "admissible" for φ (i.e., it has the syntactic shape of a well-formed argument having φ as one of its conclusions) and in addition φ is implicitly believed/known (in the sense of Kripke's relational semantics). However, as we argued in a previous paper [14], the Fitting semantics is "Gettierizable," by which we mean that it is vulnerable to counterexamples of the type given by Gettier in his celebrated paper [32]. Moreover, we argued in [14] that this semantics is prone to other more basic problems: it leads to the counterintuitive conclusion that an agent must accept (believe in the legitimacy of) a body of evidence t supporting an assertion φ only because she believes the conclusion φ, even if in fact her belief in φ has nothing to do with the evidence t.[1] In the same paper, we sketched a proposal for an alternative semantics that could address this problem.

In this paper, we build on the work in [14] by proposing a new solution to the problem of the "Gettierizability" of Fitting semantics, a solution based on introducing a notion of *actual availability of "conclusive" (or "good") evidence*. A conclusive body of evidence is one that is fully reliable, in the sense that it is truthful whenever it is available, and moreover all of its component pieces of evidence are similarly reliable. Intuitively, we say that the body of evidence t is "explicit good evidence" (actually available to the agent) at a world w (and write that Et holds at w) if the agent has "constructed" (i.e., computed, observed, etc.) t and if in addition t is indeed conclusive evidence at world w. Despite what the name may suggest, an agent can still be mislead about whether Et holds: she might believe some constructed evidence to be conclusive, while in fact it is not. And conversely, t can be explicit good evidence without being actually accepted by the agent: although the evidence t is available to her (say, because she has observed or constructed t), she does not believe it to be fully reliable (though in fact it is). So whether evidence is accepted (i.e., believed, or "known," to be conclusive/legitimate) is independent of whether that evidence is in fact legitimate.

We present here a static logical account for reasoning about the notions of conclusive evidence, justifiable belief (and conditional belief), defeasible knowledge, and ("hard") information. In particular, while an agent may have an implicit belief, which may be justifiable in principle by some legitimate piece of evidence that is implicitly accepted by the agent, it need not be the case that this belief is explicit: the required evidence might not be currently available to the agent (due to her computational limits or to a lack of time). Only evidence that is available to the agent (via conscious observation, explicit logical construction, or actual computation) can serve as explicit justification for the agent's beliefs. When evidence is both available to the agent and accepted as legitimate by her, then any assertion supported by it is *explicitly believed*: the evidence provides the agent with an *explicit justification* for this belief. If in addition the evidence is also known to be conclusive, then any assertion supported by it is *explicitly known*: the evidence provides the agent with an *explicit conclusive justification* for this piece of knowledge.

It will therefore be useful to distinguish between the *implicit* notions of belief, conditional belief, defeasible knowledge, and ("hard") information and the *explicit* statements that are supported by specific pieces of evidence that say why the statement in question holds. As we will see, the

[1] More precisely, if the agent implicitly believes φ in the sense of Kripke's semantics, then *any* evidence t that is admissible for φ justifies the belief. In particular, the semantics does not allow us to have two different pieces of evidence t and t', both admissible for a believed formula φ, and yet just one of these pieces of evidence is the justification. Further, an admissible t for a believed formula φ always provides a justification, even if in fact the agent does not believe some of the other assertions supported either by t or by one of its constituent pieces of evidence!

implicit notions are closed under logical consequence and therefore suffer from the *problem of logical omniscience*, wherein the notion in question simply attributes too much cognitive power to the agent, having her, for example, believe *all* logical consequences of her beliefs. The explicit notions, on the other hand, need not be closed under logical consequences, though generally closure will hold to some reasonable finite degree. As a result, evidence-based justified belief provides us with an account of agent reasoning that addresses logical omniscience. (We discuss this later in §6.)

The work in this paper is a new development of the "static" sub-system of the logic of dynamic evidence presented in [14]. The main difference is that the logic studied here is based on the notion of conclusive evidence and, for simplicity, we restrict here to the *static* case (i.e., we do not address the question of how evidence and explicit evidence-based cognitive notions develop over time as a consequence of evidence-presenting informational actions such as public announcements, private messages, radical upgrades, or the like). Nevertheless, we are still able to talk about *potential* dynamics: conditional explicit beliefs are a way to "pre-encode" evidential changes.

The logics of the implicit cognitive notions studied here come from work by two of the authors [15] that studied these and related notions in detail. Here we take that work and, following in the spirit of the Justification Logic approach to evidence [10] and building on the setting of justifiable belief and explicitly constructable evidence from our previous work [14], we take the first steps towards developing a new framework for formal epistemology: a *theory of explicitly justified knowledge based on the actual availability of conclusive evidence.*

2 Syntax

Definition 2.1 (Language)**.** Given a set Φ of atomic sentences, the language $\mathscr{L} := (\mathscr{T}, \mathscr{F})$ consists of the set $\mathscr{T}$ of *evidence terms* t and the set $\mathscr{F}$ of *propositional formulas* (sentences) φ defined by the following double recursion:

$$\begin{array}{lcl} \varphi & ::= & \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid t \gg \varphi \mid Et \mid K\varphi \mid \Box\varphi \quad \text{with } p \in \Phi \\ t & ::= & c_\varphi \mid t \cdot t \mid t + t \end{array}$$

An *atomic term* is one of the form c_φ, also called a *certificate for* φ. A *compound term* is one that is non-atomic. Notation: let $\hat{K}$ denote $\neg K \neg$, let $\Diamond$ denote $\neg\Box\neg$, and let $\bot$ denote $\neg\top$.

The set $\mathsf{sub}(\varphi)$ of *subformulas* of a formula φ is defined by induction on the construction of φ as follows: $\mathsf{sub}(\top) := \{\top\}$, $\mathsf{sub}(p) := \{p\}$, $\mathsf{sub}(\neg\theta) := \{\neg\theta\} \cup \mathsf{sub}(\theta)$, $\mathsf{sub}(\theta \wedge \theta') := \{\theta \wedge \theta'\} \cup \mathsf{sub}(\theta) \cup \mathsf{sub}(\theta')$, $\mathsf{sub}(Et) := \{Et\}$, $\mathsf{sub}(t \gg \theta) := \{t \gg \theta\} \cup \{\theta\}$, $\mathsf{sub}(K\theta) := \{K\theta\} \cup \mathsf{sub}(\theta)$, and $\mathsf{sub}(\Box\theta) := \{\Box\theta\} \cup \mathsf{sub}(\theta)$. The set $\mathsf{sub}(t)$ of *subterms* of a term t is defined by induction on the construction of t as follows: $\mathsf{sub}(c_\varphi) := \{c_\varphi\}$, $\mathsf{sub}(s \cdot u) := \{s \cdot u\} \cup \mathsf{sub}(s) \cup \mathsf{sub}(u)$, and $\mathsf{sub}(s + u) := \{s + u\} \cup \mathsf{sub}(s) \cup \mathsf{sub}(u)$.

We think of terms as pieces of evidence that justify various formulas. The term c_φ is an *evidential certificate*: a "canonical" piece of evidence in support of formula φ. The term $t \cdot s$ is a compound piece of evidence obtained by combining the two pieces of evidence t and s using Modus Ponens. The term $t + s$ is a body of evidence that aggregates (without performing logical inference) all the evidence provided by t and by s, so that $t + s$ supports all those things supported by one or more of t and s.

The formula $t \gg \varphi$ says that *evidence* t *is admissible for* φ, which intuitively means that: t has the formal structure of a well-formed argument, and, moreover, the structure of an argument in

favor of φ. In other words, t "looks" like a possible justification for φ. Admissibility is a purely *formal* notion: it simply checks that t has the correct syntactic shape to qualify as an argument for φ. This says nothing about the validity of this possible argument, nor about its acceptance or its persuasive power. In fact, it very well might be the case that t is formally admissible for φ while in fact φ is neither true nor is it believed: the truth (or acceptance) of the conclusion φ depends not only on the formal correctness of the argument t but also on the truth (or acceptance) of its premises. So essentially our formula $t \gg \varphi$ expresses a purely syntactic meta-relation between terms and sentences. We define this meta-relation inductively on the syntax as follows:

Definition 2.2. *Admissibility* is the smallest binary relation $\gg \subseteq \mathscr{T} \times \mathscr{F}$ satisfying the following conditions:

(a) $c_\varphi \gg \varphi$;

(b) if $t \gg (\psi \Rightarrow \varphi)$ and $s \gg \psi$, then $(t \cdot s) \gg \varphi$; and

(c) if $t \gg \varphi$ and $s \gg \psi$, then $(t + s) \gg \varphi$ and $(t + s) \gg \psi$.[2]

Though we use the same word ("admissibility") and same symbol ("$\gg$") both for the just-defined syntactic meta-relation and for a symbol in our formal language (that internalizes this meta-relation in our syntax), this ambiguity is innocuous: it will be obvious from context which is which.

Note that our definition of admissibility differs in two ways from the definition that is usually adopted in the Justification Logic literature, which comes by way of the Fitting semantics [10, 29]. First, our admissibility condition for $t + s$ differs from the usual one, in that we require *both t and s to be admissible for some sentences* in order for $t + s$ to be admissible for any sentence (in which case $t + s$ is as usual admissible for any sentence for which either t or s is admissible). This leads to our condition (c) above. In contrast, the standard Fitting condition is: (c′) if $t \gg \varphi$ or $s \gg \varphi$, then $(t + s) \gg \varphi$. The reason we choose (c) over (c′) is that we understand admissibility of t as implying that *t is a well-shaped argument*: for this to be the case, all its components should also be well-shaped arguments (i.e., admissible for some sentence). The second major difference from Fitting's notion is that we take admissibility as the *smallest* binary relation satisfying the conditions (a)–(c), whereas the traditional approach in Justification Logic (coming by way of the Fitting semantics [10, 29]) would have us take *any* binary relation satisfying the conditions (a), (b), and (c′).[3] One key consequence of our departure from the traditional approach is that every compound term t that is admissible for a formula φ is admissible for that formula in virtue of a "derivation" of $t \gg \varphi$ from a finite number of assumptions of the form $c_\psi \gg \psi$ using the conditions (b) and (c) of Definition 2.2 as "rules." It is in this sense that our notion of admissibility ensures that there is a "chain of evidence" linking the support a term t gives to a formula φ with a stepwise derivation from certificates. In the Fitting semantics, this is not always the case: there are Fitting models in which a compound term may be admissible for a formula φ without the proper subterms of t being admissible for *any* formula. This breaks the "chain of evidence," in that the support given by t to

[2] Note that principle (c) also encompasses the following: if $t \gg \varphi$ and $s \gg \psi$, then $(s + t) \gg \varphi$ and $(s + t) \gg \psi$. Since $\gg$ is the smallest binary relation satisfying (a)–(c), it follows that $(t + s) \gg \varphi$ iff $(s + t) \gg \varphi$ for each formula φ. The operator $+$ is in this sense a "commutative aggregator" of evidence.

[3] Actually, following the Fitting semantics, we would take admissibility to be a relation that satisfies the conditions (and so contains the smallest relation satisfying the conditions) but that otherwise varies from model to model. We do not take this approach here.

φ has nothing to do with the support given by the proper subterms of t to any other formulas. In our setting, such a possibility is expressly forbidden: first, if t supports φ, then, according to our rules, every proper subterm of t supports (and hence is admissible for) some formula; and second, if t supports φ, then there exists a stepwise "evidential chain" that begins with a finite number of certificates, ends with t itself, and includes at each intermediary step only subterms of t supporting formulas that are relevant for the derivation of φ.

Let us explain in more detail our admissibility conditions. We have $c_\varphi \gg \varphi$ because certificates by definition are meant to certify: so c_φ does have the right syntactic shape of a justification for φ (even though this justification might not be valid or accepted). $t{\cdot}s$ denotes an argument constructed from t and s by applying a Modus Ponens step, so that $t \cdot s$ is a well-formed argument for φ iff there exists a sentence ψ such that t is a well-formed argument for $\psi \Rightarrow \varphi$ and s is a well-formed argument for ψ. Finally, $t + s$ is obtained by aggregating two well-formed arguments.

Definition 2.3 (Content)**.** For every term $t \in \mathscr{T}$, we define the *content* con_t *of* t as the set of all the formulas for which t is admissible evidence:

$$\mathsf{con}_t := \{\theta \in \mathscr{F} \mid t \gg \theta\} \ .$$

When convenient, we will write con_t as $\mathsf{con}(t)$.

Definition 2.4. $\mathscr{T}^e := \{t \in \mathscr{T} \mid \mathsf{con}_t \neq \emptyset\}$ is the *set of admissible terms.*

Lemma 2.5 (Computability of Admissibility)**.** The map $t \mapsto \mathsf{con}_t$ of type $\mathscr{T} \to \wp(\mathscr{F})$ is computable, and con_t is finite for every $t \in \mathscr{T}$.

We note that $\mathscr{T}^e \neq \mathscr{T}$ since $\mathsf{con}(c_\top \cdot c_{\top\Rightarrow\top}) = \emptyset$.[4] If $t \in \mathscr{T} - \mathscr{T}^e$, we have $\bigwedge \mathsf{con}_t = \top$. Otherwise, if $t \in \mathscr{T}^e$, then Lemma 2.5 implies $\bigwedge \mathsf{con}_t$ is a finite conjunction and hence a formula. The formula $\bigwedge \mathsf{con}_t$ characterizes the content of term t. The content may be thought of as the information that the evidence term can reasonably be construed to support.

A key concept in our theory is the *explicit availability of good evidence Et* ("explicit goodness" for short). This means that t is a truth-based body of evidence that is explicitly available to the agent. In other words, "explicit goodness" Et comprises two conditions:

(i) the agent has actually observed (via correct, sound observations) all the basic pieces of evidence used as premises in the argument t (so that these premises are true and available to the agent); and

(ii) the agent has actually constructed the argument t.

Since Et is meant to indicate legitimacy of use, we will require in the semantics that every explicitly good evidence term s has true content $\bigwedge \mathsf{con}_s$, a condition we call "evidential goodness."

It is equally important to note what is *not* implied by our notion of explicit goodness. Indeed, Et may hold, while we may also have: (i′) the agent may not believe her own observations (of the above-mentioned basic pieces of evidence) to be reliable, or, even if she believes them to be so, she may not know them to be so (but instead has doubts concerning her observations); (ii′) the agent may have used the correct premises in an incorrect way by constructing a poorly shaped argument t (that may contain parts that are not formally correct and hence are not admissible for any sentence).

[4]The Modus Ponens operation $\cdot$ is not commutative: $\{\top\} = \mathsf{con}(c_{\top\Rightarrow\top} \cdot c_\top) \neq \mathsf{con}(c_\top \cdot c_{\top\Rightarrow\top}) = \emptyset$.

We use the formula $K\varphi$ to say that φ *follows from the "hard" information the agent has received*, so that the dual formula $\hat{K}\varphi$ says that φ *is consistent with the "hard" information the agent has received*. One may consider K as a form of knowledge, though not the kind of knowledge that is usually possessed by actual, real-life agents. This sort of knowledge can be thought of as "implicit" (since it does not have anything to do with the agent's mental state, but only with what implicitly follows from her information). It is also unrevisable, as well as "infallible": it cannot be defeated in any way and it cannot fail to be true. The best term for K is thus "information" (or "hard information") rather than "knowledge."

In contrast, the formula $\Box\varphi$ captures a more realistic (though *still implicit* or potential) notion of knowledge: "the agent defeasibly knows φ." This is an embodiment of the so-called Defeasibility Theory of Knowledge, championed mainly by Lehrer [47, 48] and Lehrer and Paxson [49]: φ is implicitly "known" in this sense if it is implicitly believed and would continue to be implicitly believed upon receipt of any further *truthful* information. So $\Box$ refers to an implicit form of "robust" belief: one that cannot be defeated by any truthful evidence.[5]

Defeasible knowledge is always taken with respect to the information the agent has received. Therefore, everything consistent with the agent's defeasible knowledge is also consistent with the information she has received: $\Diamond\varphi \Rightarrow \hat{K}\varphi$.

As already mentioned, $\Box$ is still *an implicit or potential kind of "knowledge"*: the agent has access to it in principle but may fail to explicitly possess it. We will later combine $\Box$ with the explicit availability of good evidence to capture *an explicit (or "conscious") type of defeasible knowledge*: $\Box^e\varphi$ means that the agent actually knows φ.

Definition 2.6. We introduce the following abbreviations:

$B\varphi$	$:=$	$\Diamond\Box\varphi$	(implicit belief)
$B(\varphi\|\psi)$	$:=$	$\hat{K}\psi \Rightarrow \hat{K}(\psi \wedge \Box(\psi \Rightarrow \varphi))$	(conditional belief)
$t:\varphi$	$:=$	$t \gg \varphi \wedge \bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta$	(implicit justification)
$t:^e \varphi$	$:=$	$Et \wedge t:\varphi$	(explicit justification)
$B^e\varphi$	$:=$	$B(Ec_\varphi)$	(explicit belief)
$B^e(\varphi\|\psi)$	$:=$	$B(Ec_\varphi\|\psi)$	(explicit conditional belief)
$\Box^e\varphi$	$:=$	$\Box(Ec_\varphi)$	(explicit defeasible knowledge)
$K^e\varphi$	$:=$	$K(Ec_\varphi)$	(explicit information)

Intuitively, $B\varphi$ means that *the agent implicitly believes* φ; i.e., that φ follows from some of the agent's beliefs. The usual Grove semantics of belief in plausibility models [15] makes $B\varphi$ equivalent to the statement that φ is true in all the most plausible worlds. Note that in our logic, B is defined as a simple abbreviation for $\Diamond\Box$. As shown later, in our intended ("standard") models, the meaning of this abbreviation matches the usual Grove semantics.[6]

[5] Such robustness may suggest that "*indefeasible* knowledge" would be a more appropriate name. But the standard terminology is justified by noticing that defeasible knowledge *can* be defeated, namely by *false* information (e.g., lies); in contrast, the "hard information" embodied by $K\varphi$ is truly indefeasible!

[6] Another way to think of the abbreviation $B\varphi := \Diamond\Box\varphi$ is the following: implicit belief $B\varphi$ is defeasible knowledge that is "good enough," by which we mean that, though the agent does not defeasibly know φ outright, it is at least consistent with her defeasible knowledge that she defeasibly knows φ.

Similar comments apply to conditional belief. Intuitively, $B(\varphi|\psi)$ says that *the agent implicitly believes φ conditional on ψ*; i.e., φ follows from the agent's beliefs about what would be the case if ψ turned out to be the case. The usual semantics of conditional belief in plausibility models [15] makes $B(\varphi|\psi)$ equivalent to the statement that φ is true in all the most plausible ψ-worlds. Once again, the meaning of conditional belief in our setting will be seen to match the usual semantics in standard models.

Among the abbreviations introduced above, there are two key higher-level evidential concepts: the twin concepts of

- *justification of φ by implicit conclusive evidence t* (written $t : \varphi$) and
- *justification of φ by explicit conclusive evidence t* (written $t :^e \varphi$).

The first, $t : \varphi$, captures the fact that t *is an entirely correct argument for* φ: not only does t have the correct shape of an argument supporting φ (i.e., we have $t \gg \varphi$), but t is also based only on correct (i.e., true) premises (i.e., we have $\bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta$). But while t may be an entirely correct argument for φ (i.e., $t : \varphi$ is true), this argument may nevertheless be unavailable to the agent either because the premises are not known to her or because she did not yet succeed in constructing this argument (due to her bounded computational-logical resources or simply due to a lack of time). If in addition to $t : \varphi$ we also have evidential goodness Et (and so in particular the argument t and its premises are actually available to the agent), then t qualifies as "explicit conclusive evidence"; i.e., we have $t :^e \varphi$.

Finally, the abbreviations $B^e\varphi$, $B^e(\varphi|\psi)$, $\Box^e\varphi$, and $K^e\varphi$ capture the agent's *explicit* doxastic/epistemic attitudes: the explicit, conscious possession of belief, conditional belief, defeasible knowledge, and hard information. For example, $B^e\varphi := B(Ec_\varphi)$ says that the agent believes she is in actual possession of the good evidence c_φ. If this is so, then, in virtue of the goodness of this evidence, she must believe that the evidential content $\bigwedge \mathsf{con}(c_\varphi) = \varphi$ is true. The agent therefore implicitly believes φ on the basis of the certificate c_φ, which she implicitly believes she actually possesses. For present purposes, this is tantamount to the agent having an explicit, conscious possession of a belief of φ. A similar discussion goes with each of $B^e(\varphi|\psi)$, $\Box^e\varphi$, and $K^e\varphi$.

3 Semantics

Definition 3.1 (Pre-model)**.** A *pre-model* $M = (W, [\![\cdot]\!], \sim, \geq, E)$ is a structure consisting of a nonempty set W of *possible worlds*, a *valuation map* $[\![\cdot]\!] : \Phi \to \wp(W)$, binary relations $\sim$ and $\geq$ on W, and a *good-evidence availability map* $E : W \to \wp(\mathscr{T})$. The components of the structure M must satisfy the following conditions:

- $\sim$ is an equivalence relation and $\geq$ is a preorder.[7]

  $w \sim v$ says, "w and v cannot be distinguished under the currently available information."

  $w \geq v$ says, "w is no more plausible than v."

  Intuitively, it is the "smaller" worlds that are more plausible: $w > v$ says that v is strictly *more* plausible than w. We define the *equi-plausibility relation* $\simeq := \geq \cap \leq$, so that $w \simeq v$ if and only if $(w \geq v) \wedge (w \leq v)$.

[7] A *preorder* is a reflexive and transitive binary relation. For a preorder $\geq$, we denote by $>$ the strict version given by $w > v := (w \geq v) \wedge (v \ngeq w)$. We denote by $\leq$ and $<$ the converse relations.

- *Local Connectedness*: $w \sim v \Rightarrow (w \geq v \vee v \geq w)$.

  Worlds consistent with the information received are always comparable in terms of plausibility.

- *Indefeasibility*: $w \geq v \Rightarrow w \sim v$.

  Worlds consistent with defeasible knowledge are consistent with the information received.

- *Trivial Evidence*: $c_\top \in E(w)$.

  The agent always has some good evidence available: she has available the certificate for the trivial tautology $\top$.

- *Certification of Evidence*: If $t \in E(w)$ and $t \gg \varphi$, then $c_\varphi \in E(w)$.

  All good available evidence is certified.

- *Subterm Closure*: If $t \cdot s \in E(w)$ or $t + s \in E(w)$, then $t, s \in E(w)$.

  If an evidential term is available as good evidence, then so are its subterms.

- *Availability of Evidence*:

  - If $t \cdot s \in E(w)$, $w \sim w'$, and $t, s \in E(w')$, then $t \cdot s \in E(w')$.
  - If $t + s \in E(w)$, $w \sim w'$, and $t, s \in E(w')$, then $t + s \in E(w')$.

    If an evidential term is actually available as good evidence, then this fact follows from the agent's information together with the goodness of its subterms.

  The last condition can be justified by our analysis of E in terms of "goodness" (or "conclusiveness") plus actual availability (i.e., constructibility, observability, or actual "possession" of the evidence). The intuition is that availability is potentially introspective: the agent has implicit information about which arguments she has constructed/observed (regardless of whether she is explicitly aware of this). So, if a composite argument of the form $t+s$ (or $t \cdot s$) is really available to her as good evidence, then its availability is implicit information, though its "goodness" might not be. But as a consequence, the agent implicitly "knows" that *if* the constituent parts t and s are good, *then* the whole argument is good.

A *pointed pre-model* is a pair (M, w) consisting of a pre-model M and a designated world w in M called the "actual world."

Definition 3.2 (Truth)**.** We now define a *satisfaction relation* $(M, w) \models \varphi$ between pointed pre-models (M, w) and formulas $\varphi \in \mathscr{F}$. We also denote $(M, w) \models \varphi$ in the more familiar way by $w \models_M \varphi$, omitting the subscript M when M is fixed.

$$
\begin{array}{lll}
w \models \top & & \\
w \models p & \text{iff} & w \in [\![p]\!] \\
w \models \neg\varphi & \text{iff} & w \not\models \varphi \\
w \models \varphi \wedge \psi & \text{iff} & w \models \varphi \text{ and } w \models \psi \\
w \models t \gg \varphi & \text{iff} & t \gg \varphi \\
w \models Et & \text{iff} & t \in E(w) \\
w \models K\varphi & \text{iff} & v \models \varphi \text{ for every } v \sim w \\
w \models \Box\varphi & \text{iff} & v \models \varphi \text{ for every } v \leq w
\end{array}
$$

We extend the valuation map $[\![\cdot]\!]$ to *all* formulas by setting $[\![\varphi]\!] := \{w \in W \mid w \models \varphi\}$.

A first observation is that, although our syntactic definitions of belief $B\varphi$ and conditional belief $B(\varphi|\psi)$ look very different, the first is semantically equivalent to a special case of the second:

Proposition 3.3. In every pre-model, we have that:

$$w \models B\varphi \quad \text{iff} \quad w \models B(\varphi|\top) \ .$$

Proof. A semantic version of this result appears in [15]; here we prove the syntactic version. Suppose $w \models B\varphi$. This means $w \models \Diamond\Box\varphi$, which implies that there is a $v \leq w$ such that

$$u \leq v \text{ implies } u \models \varphi \ . \tag{1}$$

We want to show that $w \models B(\varphi|\top)$; that is, that

$$w \models \hat{K}\top \Rightarrow \hat{K}(\top \wedge \Box(\top \Rightarrow \varphi)) \ . \tag{2}$$

Since $w \sim w$ and $w \models \top$, we have $w \models \hat{K}\top$ and hence, to complete the argument, it suffices by (2) for us to show that

$$w \models \hat{K}(\top \wedge \Box(\top \Rightarrow \varphi)) \ . \tag{3}$$

By (1), it follows that $v \models \Box\varphi$ and hence (as is readily verified) that $v \models \top \wedge \Box(\top \Rightarrow \varphi)$. Applying the latter to the fact that $v \leq w$ implies $v \sim w$, we have (3).

For the converse direction, suppose $w \models B(\varphi|\top)$, which means that (2). Since $w \models \hat{K}\top$, it follows that (3). The latter means that there is a $v \sim w$ such that $v \models \top \wedge \Box(\top \Rightarrow \varphi)$, which itself implies (1). There are two cases to consider: $v \leq w$ and $v \not\leq w$. Now if $v \leq w$, then it follows by (1) that $w \models \Diamond\Box\varphi$ (i.e., that $w \models B\varphi$), and we are done. So suppose that $v \not\leq w$. Since $v \sim w$, we have $w < v$. But every $u \leq w$ then satisfies $u \leq v$ and hence we have by (1) that $w \models \Box\varphi$. Since $w \leq w$, it again follows that $w \models \Diamond\Box\varphi$ (i.e., that $w \models B\varphi$). □

The second observation is that, if we restrict our pre-models to the ones matching Grove's standard assumptions about plausibility [34], our notion of (conditional) belief is semantically equivalent to the Grove's well-known (and very natural) definition of (conditional) belief $B(\varphi|\psi)$ as "truth in the most plausible ψ-worlds (that are indistinguishable from the actual world)":

Definition 3.4. *The Best Worlds Assumption* is the statement that for every nonempty set $P \subseteq W$ of informationally indistinguishable worlds (i.e., such that $w \sim w'$ for all $w, w' \in P$), the set

$$\min P := \{w \in P \mid w \leq w' \text{ for all } w' \in P\}$$

(consisting of the most plausible worlds in P) is also non-empty.

Proposition 3.5. Let M be a pre-model satisfying the Best Worlds Assumption. Then we have that:

$$w \models B(\varphi|\psi) \quad \text{iff} \quad \min\{w' \in [\![\psi]\!] \mid w' \sim w\} \subseteq [\![\varphi]\!] \ .$$

Proof. A semantic version of this result appears in [15]; here we prove the syntactic version. Define the set

$$[w|\psi] := \{w' \in [\![\psi]\!] \mid w' \sim w\} \ .$$

Assume $w \models B(\varphi|\psi)$. This means

$$w \models \hat{K}\psi \Rightarrow \hat{K}(\psi \wedge \Box(\psi \Rightarrow \varphi)) \ . \tag{4}$$

We want to show that $\min[w|\psi] \subseteq [\![\varphi]\!]$. We may assume that $\min[w|\psi] \neq \emptyset$, for otherwise the desired subset relationship follows immediately. So choose an arbitrary $v \in \min[w|\psi]$. That v is a member of this set means

$$v \sim w \ , \tag{5}$$

$$v \models \psi \ , \text{ and} \tag{6}$$

$$u \sim w \text{ and } u \models \psi \text{ implies } v \leq u \ . \tag{7}$$

Since $v \in \min[w|\psi]$ was chosen arbitrary, to show $\min[w|\psi] \subseteq [\![\varphi]\!]$, it suffices for us to show that $v \models \varphi$. Proceeding, it follows by (5) and (6) that $w \models \hat{K}\psi$. Therefore, it follows by (4) that there is a $u \sim w$ such that

$$u \models \psi \wedge \Box(\psi \Rightarrow \varphi) \ .$$

Since $u \sim w$ and $u \models \psi$, it follows by (7) that $v \leq u$. Since $u \models \Box(\psi \Rightarrow \varphi)$ and $v \leq u$, it follows that $v \models \psi \Rightarrow \varphi$ and therefore by (6) that $v \models \varphi$.

For the converse direction, we assume that

$$\min[w|\psi] \subseteq [\![\varphi]\!] \tag{8}$$

and prove that $w \models B(\varphi|\psi)$; that is, we wish to prove that (4). So suppose $w \models \hat{K}\psi$. This means that there is a $w' \sim w$ such that $w' \models \psi$. Hence $[w|\psi] \neq \emptyset$. Applying the Best Worlds Assumption, there is a $v \in \min[w|\psi]$. But then we have (5), (6), and (7). Now if $u \leq v$ satisfies $u \models \psi$, then $u \sim v \sim w$, so $u \sim w$, and hence $u \in \min[w|\psi]$, which implies $u \models \varphi$ by (8). That is, we have shown that $v \models \Box(\psi \Rightarrow \varphi)$. Combining this with (5) and (6), we have shown that $w \models \hat{K}(\psi \wedge \Box(\psi \Rightarrow \varphi))$. Since this was shown under the assumption that $w \models \hat{K}\psi$, our overall argument proves (4), which is it what it means to have $w \models B(\varphi|\psi)$. □

Some authors (including Grove himself [34]) assumed a weaker version of the Best Worlds Assumption, covering only the sets P that are *definable* by some sentence ψ in their language (i.e., $P = [\![\psi]\!]$ for some formula ψ). Indeed, it is easy to see that the Best Worlds Assumption can be replaced in Proposition 3.5 by this weaker condition [15]. However, in this paper, we will consider an even stronger (but mathematically simpler) condition called *standardness*:

Definition 3.6 (Standard Pre-model)**.** A pre-model $M = (W, [\![\cdot]\!], \sim, \geq, E)$ is said to be *standard* if the strict converse-plausibility relation $<$ is well-founded. This means that there are no infinite chains $w_0 > w_1 > w_2 > \cdots$ of more and more plausible worlds.

Note that *every standard pre-model satisfies the Best Worlds Assumption.*

Definition 3.7 (Model, Evidential Goodness, Validity)**.** A *model* is a pre-model satisfying the following property.

- *Evidential Goodness*: If $c_\varphi \in E(w)$, then $w \models \varphi$.

  This says that basic pieces of evidence (i.e., certificates) that are available as good evidence are indeed "good" (or "conclusive"): what they certify is true.

Validity $\models \varphi$ means that $(M, w) \models \varphi$ for every *pointed standard model* (M, w).

Proposition 3.8 (Good Evidence)**.** Every term t that is available as good evidence is indeed "good" (i.e., its content is true); that is,

$$\models Et \Rightarrow \bigwedge \mathsf{con}_t \ .$$

In particular, $Ec_\varphi \Rightarrow \varphi$ is valid.

Proof. By induction on the construction of t. In the base case, $t = c_\varphi$ and $\bigwedge \mathsf{con}(c_\varphi) = \varphi$. So we want to show that $\models Ec_\varphi \Rightarrow \varphi$. Proceeding, let M be a model for which $w \models_M Ec_\varphi$. This means $c_\varphi \in E(w)$. Since M is a model, it follows by Evidential Goodness that $w \models_M \varphi$, as desired.

Inductive step $t = u \cdot v$. We have

$$\mathsf{con}(u \cdot v) = \{\psi \mid \exists \varphi : u \gg (\varphi \Rightarrow \psi) \text{ and } v \gg \varphi\} \ .$$

If $\mathsf{con}(u \cdot v) = \emptyset$, so that $\bigwedge \mathsf{con}(u \cdot v) = \top$, we have $\models E(u \cdot v) \Rightarrow \bigwedge \mathsf{con}(u \cdot v)$ immediately. Therefore, let us consider the case $\mathsf{con}(u \cdot v) \neq \emptyset$. It suffices for us to show that $\models E(u \cdot v) \Rightarrow \psi$ for an arbitrarily chosen $\psi \in \mathsf{con}(u \cdot v)$. Proceeding with such a choice of ψ along with a φ (guaranteed by the membership $\psi \in \mathsf{con}(u \cdot v)$) satisfying $u \gg (\varphi \Rightarrow \psi)$ and $v \gg \varphi$, let M be a model for which $w \models_M E(u \cdot v)$. This means $u \cdot v \in E(w)$, from which it follows by Subterm Closure that $u, v \in E(w)$. By Certification of Evidence, we have $c_{\varphi \Rightarrow \psi}, c_\varphi \in E(w)$. Since M is a model, it follows by Evidential Goodness that $w \models_M \varphi \Rightarrow \psi$ and $w \models_M \varphi$, and hence $w \models_M \psi$, as desired.

Inductive step $t = u + v$. We have

$$\begin{aligned} \mathsf{con}(u + v) \quad = \quad & \{\varphi \mid \exists \psi : u \gg \varphi \text{ and } v \gg \psi\} \cup \\ & \{\varphi \mid \exists \psi : v \gg \varphi \text{ and } u \gg \psi\} \ . \end{aligned}$$

If $\mathsf{con}(u + v) = \emptyset$, so that $\bigwedge \mathsf{con}(u + v) = \top$, the result $\models E(u + v) \Rightarrow \bigwedge \mathsf{con}(u + v)$ follows immediately. So assume $\mathsf{con}(u + v) \neq \emptyset$ and choose an arbitrary $\varphi \in \mathsf{con}(u + v)$. We wish to show $\models E(u + v) \Rightarrow \varphi$. So let M be a model satisfying $w \models_M E(u + v)$, which means $u + v \in E(w)$. By Subterm Closure, we have $u, v \in E(w)$. Since $\varphi \in \mathsf{con}(u + v)$, either we have both $u \gg \varphi$ and $v \gg \psi$ for some ψ or else we have both $v \gg \varphi$ and $u \gg \psi$ for some ψ. In either case, it follows by $u, v \in E(w)$ and Certification of Evidence that $c_\varphi \in E(w)$. Since M is a model, it follows by Evidential Goodness that $w \models_M \varphi$, as desired. □

As suggested by our definition of validity (Definition 3.7), *standard models are our intended models*. All the other notions of (pre-)model introduced above are only auxiliary concepts (though of course all the equivalences proved above for these more general notions of model will still hold for standard models). So it is important to point out that our notion of validity is non-trivial:

Proposition 3.9 (Non-triviality)**.** There exists a standard model.

Proof. Define $M = (\{w\}, [\![\cdot]\!], \sim, \geq, E)$ by $w \sim w$, $w \geq w$, $[\![r]\!] := \{w\}$ for all $r \in \Phi$, and $E(w) := \{c_\top\}$. Clearly, $\sim$ is an equivalence and $\geq$ is a preorder. Also, it is not difficult to check that M satisfies each of Local Connectedness, Indefeasibility, Trivial Evidence, Certification of Evidence, Subterm Closure, Availability of Evidence, Standardness, and Evidential Goodness. □

3.1 Defeasible Knowledge

Note our definition of $\Box$ as "truth at all the worlds that are at least as plausible as the actual world" (Definition 3.2). This formalization of defeasible knowledge in terms of plausibility orders is due to Stalnaker [56, 57]. This notion was further studied in [15] by two of the authors, who provided a complete axiomatization of the logic $K\Box$ of defeasible knowledge and hard information, and showed that other doxastic notions (such as belief, conditional belief, and strong belief) are definable in this logic. Our axiomatic system in this paper is an extension of that axiomatization.

At a first sight, our formal definition of $\Box$ looks very different from our above informal explanation of defeasible knowledge as "belief that cannot be defeated by any true information." However, it is easy to see that the two are equivalent:

Proposition 3.10. In every pre-model, we have that:

$$w \models \Box\varphi \quad \text{iff} \quad w \models B(\varphi|\psi) \text{ for all } \psi \text{ such that } w \models \psi \ .$$

Proof. A semantic version of this result appears in [15]; here we prove the syntactic version. Assume $w \models \Box\varphi$. Choose an arbitrary ψ satisfying $w \models \psi$. We wish to show that $w \models B(\varphi|\psi)$, which means

$$w \models \hat{K}\psi \Rightarrow \hat{K}(\psi \wedge \Box(\psi \Rightarrow \varphi)) \ .$$

Since $w \models \psi$ and $w \sim w$, it suffices to show that $w \models \Box(\psi \Rightarrow \varphi)$. But $w \models \Box\varphi$ implies $w \models \Box(\psi \Rightarrow \varphi)$, and so the result follows.

For the converse direction, assume that $w \models B(\varphi|\psi)$ for all formulas ψ such that $w \models \psi$. We wish to show that $w \models \Box\varphi$. Toward a contradiction, suppose that $w \not\models \Box\varphi$, so $w \models \neg\Box\varphi$. By our initial assumption, it follows that $w \models B(\varphi|\neg\Box\varphi)$. That is,

$$w \models \hat{K}\neg\Box\varphi \Rightarrow \hat{K}(\neg\Box\varphi \wedge \Box(\neg\Box\varphi \Rightarrow \varphi)) \ .$$

Since $w \sim w$ and $w \models \neg\Box\varphi$, it follows that $w \models \hat{K}\neg\Box\varphi$ and hence that

$$w \models \hat{K}(\neg\Box\varphi \wedge \Box(\neg\Box\varphi \Rightarrow \varphi)) \ .$$

This means that there is a $v \sim w$ such that

$$v \models \neg\Box\varphi \wedge \Box(\neg\Box\varphi \Rightarrow \varphi) \ . \tag{9}$$

But this implies that there is a $u \leq v$ such that $u \models \neg\varphi$. Therefore, since $u \leq u$, it follows that $u \models \neg\Box\varphi$. But $u \leq v$ and so it follows by (9) that $u \models \neg\Box\varphi \Rightarrow \varphi$, from which it follows by $u \models \neg\Box\varphi$ that $u \models \varphi$. Hence $u \models \neg\varphi$ and $u \models \varphi$, a contradiction. We conclude that our assumption $w \not\models \Box\varphi$ must have been incorrect and therefore that we in fact must have $w \models \Box\varphi$, as desired. $\Box$

We can also prove an *explicit* analogue of Proposition 3.10:

Proposition 3.11. In every pre-model, we have that:

$$w \models \Box^e\varphi \quad \text{iff} \quad w \models B^e(\varphi|\psi) \text{ for all } \psi \text{ such that } w \models \psi \ .$$

Proof. By Definition 2.6, we have $w \models \Box^e\varphi$ iff $w \models \Box(Ec_\varphi)$. By Proposition 3.10, this is equivalent to the statement that $w \models B(Ec_\varphi|\psi)$ holds for all ψ such that $w \models \psi$. So the desired equivalence follows because $B^e(\varphi|\psi) := B(Ec_\varphi|\psi)$. $\Box$

Proposition 3.11 shows that $\Box^e$ really captures the notion of *explicit defeasible knowledge*, as intended.

3.2 Relationship to Certain Justification Logic Principles

Table 1 lists a number of validities and non-validities. We divide this table into a number of "blocks" using horizontal lines. The first block in the table concerns the *Application* axiom from Justification Logic [10]:

$$t : (\varphi \Rightarrow \psi) \Rightarrow (s : \varphi \Rightarrow (t \cdot s) : \psi) \ .$$

In our setting, Application is valid only in its implicit justification form (using formulas $u : \chi$); the explicit justification form (using formulas $u :^e \chi$) is not valid. The second validity in the first block of the table tells us why this is so: in order for us to have $t \cdot s$ as explicit justification for ψ, not only must we have that s is explicit justification for some φ and that t is explicit justification for $\varphi \Rightarrow \psi$, but we must also have that the combined evidence $t \cdot s$ is actually available to the agent. This makes sense: the explicit justifications t and s are also implicit justifications for their respective formulas and implicit justification is closed under Modus Ponens via our operation $t \cdot s$; but to transform the implicit justification $t \cdot s$ for ψ into an explicit one, the agent must have evidence $t \cdot s$ actually available to her.

The second block in Table 1 concerns the *Sum* axiom from Justification Logic [10], $t : \varphi \vee s : \varphi \Rightarrow (t + s) : \varphi$, which is often stated as two axioms:

$$t : \varphi \Rightarrow (t + s) : \varphi \quad \text{and} \quad s : \varphi \Rightarrow (t + s) : \varphi \ .$$

Let us call the two axioms *Left Sum* and *Right Sum*, respectively. Our version of Left Sum appears as the first validity in the second block of Table 1. Since we require that the implicit justification $t + s$ of φ satisfy the property that each of its components t and s is a well-shaped argument (i.e., is admissible for some sentence), our version of Left Sum requires an additional condition in the antecedent. A similar comment holds for our version of Right Sum (the third validity in the second block). We note that the explicit versions of Left and Right Sum (the non-validities in the second block of Table 1) fail for the same reason the explicit version of Application failed: closure of justified evidence under term-combining operators ($+$ and $\cdot$) requires that the combined evidence $t + s$ (or $t \cdot s$) actually be available to the agent. The correct version of explicit Left and Right Sum are stated as the second and fourth validities in the second block of Table 1.

The third block in Table 1 concerns the *Positive Introspection* axiom (sometimes called *Checker*) of Justification Logic [10]:

$$t : \varphi \Rightarrow !t : (t : \varphi) \ .$$

Note that our language does not contain the term-forming operator !. Roughly speaking, in Justification Logic the role of the evidence $!t$ is to "check" whether t is indeed a justification of φ. Though we do not have this evidence operator in our language, we can simulate it using the term $!(t, \varphi)$ defined by the abbreviation in Table 1. Note that our "checker" term $!(t, \varphi)$, unlike the original $!t$ from Justification Logic, requires the formula φ that we are checking as a parameter. Our version of Positive Introspection appears as the first validity in the third block of Table 1. As for Application and Sum, the explicit version of Positive Introspection (the non-validity in the third block) is not valid; the valid version (the second validity in the third block) requires that the agent actually have the "explicit checker" $!^e(t, \varphi)$ available to her.

The fourth block in Table 1 concerns the *Negative Introspection* axiom of Justification Logic [10]:

$$\neg t : \varphi \Rightarrow ?t : (\neg t : \varphi) \ .$$

Validities	**Non-validities**
$t : (\varphi \Rightarrow \psi) \Rightarrow (s : \varphi \Rightarrow (t \cdot s) : \psi)$	$t :^e (\varphi \Rightarrow \psi) \Rightarrow (s :^e \varphi \Rightarrow (t \cdot s) :^e \psi)$
$t :^e (\varphi \Rightarrow \psi) \Rightarrow (s :^e \varphi \wedge E(t \cdot s) \Rightarrow (t \cdot s) :^e \psi)$	
$t : \varphi \wedge s : \psi \Rightarrow (t + s) : \varphi$	$t :^e \varphi \wedge s :^e \psi \Rightarrow (t + s) :^e \varphi$
$t :^e \varphi \wedge s :^e \psi \wedge E(t + s) \Rightarrow (t + s) :^e \varphi$	
$s : \varphi \wedge t : \psi \Rightarrow (t + s) : \varphi$	$s :^e \varphi \wedge t :^e \psi \Rightarrow (t + s) :^e \varphi$
$s :^e \varphi \wedge t :^e \psi \wedge E(t + s) \Rightarrow (t + s) :^e \varphi$	
$t : \varphi \Rightarrow !(t, \varphi) : (t : \varphi)$	$t :^e \varphi \Rightarrow !^e(t, \varphi) :^e (t :^e \varphi)$
$t :^e \varphi \wedge E!^e(t, \varphi) \Rightarrow !^e(t, \varphi) :^e (t :^e \varphi)$	
$\neg t : \varphi \Rightarrow ?(t, \varphi) : (\neg t : \varphi)$	$\neg t :^e \varphi \Rightarrow ?^e(t, \varphi) :^e (\neg t :^e \varphi)$
$\neg t :^e \varphi \wedge E?^e(t, \varphi) \Rightarrow ?^e(t, \varphi) :^e (\neg t :^e \varphi)$	
$t : \varphi \Rightarrow \varphi$	
$t :^e \varphi \Rightarrow \varphi \wedge t : \varphi \wedge Ec_\varphi$	
$B(t :^e \varphi) \Rightarrow B^e\varphi$	
$\Box(t :^e \varphi) \Rightarrow \Box^e\varphi$	
$K(t :^e \varphi) \Rightarrow K^e\varphi$	
$B^e\varphi \Rightarrow B\varphi$	
$\Box^e\varphi \Rightarrow \Box\varphi$	
$K^e\varphi \Rightarrow K\varphi$	

Abbreviations: $!(t, \varphi) := c_{t : \varphi}$ $\quad !^e(t, \varphi) := c_{t :^e \varphi}$

$?(t, \varphi) := c_{\neg t : \varphi}$ $\quad ?^e(t, \varphi) := c_{\neg t :^e \varphi}$

Table 1: Some validities and non-validities

As for the "(positive) checker" operator !, our language does not contain the "negative checker" operator ?. Roughly speaking, in Justification Logic the role of the evidence $?t$ is to validate the failure of term t to justify φ. As for the positive checker, we can simulate the negative checker operator using the term $?(t, \varphi)$ defined by abbreviation in Table 1. We note that our negative checker $?(t, \varphi)$ also differs from its original in that ours requires the formula φ as a parameter. Our version of Negative Introspection appears as the first validity in the fourth block of Table 1. As for Application, Sum, and Positive Introspection, the explicit version of Negative Introspection (the non-validity in the fourth block) is not valid; the valid version (the second validity in the fourth block) requires that the agent actually have the "explicit negative checker" $?^e(t, \varphi)$ available to her.

The fifth block in Table 1 concerns the *Factivity* axiom of Justification Logic [10]:

$$t : \varphi \Rightarrow \varphi \ .$$

Both the implicit and explicit forms of this axiom are valid in our setting, which follows our original intention: evidence for φ is always *good* (or *conclusive*). Further, $t :^e \varphi$ implies not only φ but also $t : \varphi$ (since explicit justification implies implicit justification) and Ec_φ (since the explicit justification t for φ is certified by the explicitly available certificate c_φ).

The validities in the sixth block of Table 1 posit natural relationships between implicit and explicit attitudes for each of the attitudes of belief, defeasible knowledge, and the possession of "hard" information. In particular, these validities tell us that if the agent has the attitude *implicitly* with respect to an explicit justification she has for φ, then she has the attitude *explicitly* with respect to φ. The validities in the seventh block of Table 1 describe a natural reverse relationship between the implicit and explicit attitudes: if the agent has the attitude *explicitly* with respect to φ, then she also has the attitude *implicitly* with respect to φ.

Proposition 3.12. Regarding Table 1: each of the claimed validities is valid, and each of the claimed non-validities is not.

Proof. We address only the validities. Verification of the non-validities is straightforward.

- $\models t : (\varphi \Rightarrow \psi) \Rightarrow (s : \varphi \Rightarrow (t \cdot s) : \psi)$ (Application)

  Assume $w \models t : (\varphi \Rightarrow \psi)$ and $w \models s : \varphi$. This means

  $$w \models t \gg (\varphi \Rightarrow \psi) \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta \quad \text{and} \tag{10}$$
  $$w \models s \gg \varphi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(s)} \theta \ . \tag{11}$$

  It follows from $w \models t \gg (\varphi \Rightarrow \psi)$ and $w \models s \gg \varphi$ that $w \models (t \cdot s) \gg \psi$. Since

  $$\mathsf{sub}(t) \cup \mathsf{sub}(s) \cup \{t \cdot s\} = \mathsf{sub}(t \cdot s) \ ,$$

  it follows that $c_\theta \in \mathsf{sub}(t) \cup \mathsf{sub}(s)$ iff $c_\theta \in \mathsf{sub}(t \cdot s)$ and hence we have by (10) and (11) that $w \models \bigwedge_{c_\theta \in \mathsf{sub}(t \cdot s)} \theta$. Conclusion: $w \models (t \cdot s) : \psi$.

- $\models t :^e (\varphi \Rightarrow \psi) \Rightarrow (s :^e \varphi \wedge E(t \cdot s) \Rightarrow (t \cdot s) :^e \psi)$ (Explicit Application)

  Assume $w \models t :^e (\varphi \Rightarrow \psi)$, $w \models s :^e \varphi$, and $w \models E(t \cdot s)$. This means $w \models Et \wedge t : (\varphi \Rightarrow \psi)$ and $w \models Es \wedge s : \varphi$. It follows by Application that $w \models (t \cdot s) : \psi$. But since $w \models E(t \cdot s)$, we have shown that $w \models (t \cdot s) :^e \psi$.

- $\models t:\varphi \wedge s:\psi \Rightarrow (t+s):\varphi$ (Left Sum)
  $\models t:^e \varphi \wedge s:^e \psi \wedge E(t+s) \Rightarrow (t+s):^e \varphi$ (Explicit Left Sum)
  $\models s:\varphi \wedge t:\psi \Rightarrow (t+s):\varphi$ (Right Sum)
  $\models s:^e \varphi \wedge t:^e \psi \wedge E(t+s) \Rightarrow (t+s):^e \varphi$ (Explicit Right Sum)

  We prove Left Sum. Assume $w \models t:\varphi \wedge s:\psi$. This means

  $$w \models t \gg \varphi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta \quad \text{and} \tag{12}$$
  $$w \models s \gg \psi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(s)} \theta \ . \tag{13}$$

  It follows from $w \models t \gg \varphi$ and $w \models s \gg \psi$ that $w \models (t+s) \gg \varphi$. Since

  $$\mathsf{sub}(t) \cup \mathsf{sub}(s) \cup \{t+s\} = \mathsf{sub}(t+s) \ ,$$

  it follows that $c_\theta \in \mathsf{sub}(t) \cup \mathsf{sub}(s)$ iff $c_\theta \in \mathsf{sub}(t+s)$ and hence we have by (12) and (13) that $w \models \bigwedge_{c_\theta \in \mathsf{sub}(t+s)} \theta$. Conclusion: $w \models (t+s):\varphi$.

  The argument for Explicit Left Sum follows by the abbreviation $u:^e \theta := Eu \wedge u:\theta$ and the result for Left Sum similar to the way in which the argument for Explicit Application followed by the same abbreviation and the result for Application. The arguments for Right Sum and Explicit Right Sum are similar.

- $\models t:\varphi \Rightarrow !(t,\varphi):(t:\varphi)$ (Positive Introspection)

  Assume $w \models t:\varphi$. Note that since $!(t,\varphi) := c_{t:\varphi}$, we have (by Definition 2.2) that $w \models !(t,\varphi) \gg (t:\varphi)$. Now, we have $c_\theta \in \mathsf{sub}(!(t,\varphi)) = \mathsf{sub}(c_{t:\varphi}) = \{c_{t:\varphi}\}$ iff $\theta = t:\varphi$. Hence $w \models \bigwedge_{c_\theta \in \mathsf{sub}(!(t,\varphi))} \theta$ iff $w \models t:\varphi$, but $w \models t:\varphi$ holds by assumption. We have therefore shown that

  $$w \models !(t,\varphi) \gg (t:\varphi) \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(!(t,\varphi))} \theta \ ,$$

  which means $w \models !(t,\varphi):(t:\varphi)$.

- $\models t:^e \varphi \wedge E!^e(t,\varphi) \Rightarrow !^e(t,\varphi):^e (t:^e \varphi)$ (Explicit Positive Introspection)

  Assume $w \models t:^e \varphi \wedge E!^e(t,\varphi)$. Since $!^e(t,\varphi) := c_{t:^e\varphi}$ and $c_\theta \in \mathsf{sub}(!^e(t,\varphi)) = \mathsf{sub}(c_{t:^e\varphi}) = \{c_{t:^e\varphi}\}$ iff $\theta = t:^e\varphi$, it follows by an argument like that for Positive Introspection that we have $w \models !^e(t,\varphi):(t:^e\varphi)$. But since we also have that $w \models E!^e(t,\varphi)$, it follows that $w \models !^e(t,\varphi):^e (t:^e\varphi)$, as desired.

- $\models \neg t:\varphi \Rightarrow ?(t,\varphi):(\neg t:\varphi)$ (Negative Introspection)

  Assume $w \models \neg t:\varphi$. Since $?(t,\varphi) := c_{\neg t:\varphi}$ and $c_\theta \in \mathsf{sub}(?(t,\varphi)) = \mathsf{sub}(c_{\neg t:\varphi}) = \{c_{\neg t:\varphi}\}$ iff $\theta = \neg t:\varphi$, it follows by an argument like that for Positive Introspection that we have $w \models ?(t,\varphi):(\neg t:\varphi)$.

- $\models \neg t:^e \varphi \wedge E?^e(t,\varphi) \Rightarrow ?^e(t,\varphi):^e (\neg t:^e \varphi)$ (Explicit Negative Introspection)

  Assume $w \models \neg t:^e \varphi \wedge E?^e(t,\varphi)$. Since $?^e(t,\varphi) := c_{\neg t:^e\varphi}$, a similar reasoning process as the one for Positive Introspection leads us to the conclusion that $w \models ?^e(t,\varphi):(\neg t:^e\varphi)$. But since we also have that $w \models E?^e(t,\varphi)$, it follows that $w \models ?^e(t,\varphi):^e (\neg t:^e\varphi)$, as desired.

- $\models t:\varphi \Rightarrow \varphi$ (Factivity)

  By induction on the construction of t. Proceeding with the base case, where $t = c_\psi$, we assume that $w \models c_\psi : \varphi$. This means $w \models (c_\psi \gg \varphi) \wedge \psi$. Since $w \models c_\psi \gg \varphi$ iff $\psi = \varphi$, the result follows. We now proceed to the induction step: we assume that Factivity holds for all strict subterms of t and prove that Factivity holds for t itself. There are two cases to consider: $t = u \cdot v$ and $t = u + v$. For the case $t = u \cdot v$, we assume $w \models (u \cdot v) : \varphi$. This means

  $$w \models (u \cdot v) \gg \varphi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(u \cdot v)} \theta \ . \tag{14}$$

  Now $w \models (u \cdot v) \gg \varphi$ iff there exists a ψ such that $w \models u \gg (\psi \Rightarrow \varphi)$ and $w \models v \gg \psi$. Further, since $\mathsf{sub}(u) \subseteq \mathsf{sub}(u \cdot v)$ and $\mathsf{sub}(v) \subseteq \mathsf{sub}(u \cdot v)$, it follows by (14) that $w \models \bigwedge_{c_\theta \in \mathsf{sub}(u)} \theta$ and $w \models \bigwedge_{c_\theta \in \mathsf{sub}(v)} \theta$. But then $w \models u : (\psi \Rightarrow \varphi)$ and $w \models v : \psi$, and hence $w \models \psi \Rightarrow \varphi$ and $w \models \psi$ by the induction hypothesis for u and for v. Hence $w \models \varphi$, which completes the argument for the case $t = u \cdot v$. What remains is the case $t = u + v$. We proceed by assuming that $w \models (u + v) : \varphi$, which means

  $$w \models (u + v) \gg \varphi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(u+v)} \theta \ . \tag{15}$$

  We have $w \models (u+v) \gg \varphi$ iff there exists a ψ such that, without loss of generality, $w \models u \gg \varphi$ and $w \models v \gg \psi$. Since $\mathsf{sub}(u) \subseteq \mathsf{sub}(u + v)$, it follows by (15) that $w \models \bigwedge_{c_\theta \in \mathsf{sub}(u)} \theta$. Hence $w \models u : \varphi$. Applying the induction hypothesis for u, we conclude that $w \models \varphi$, as desired.

- $\models t :^e \varphi \Rightarrow \varphi \wedge t : \varphi \wedge Ec_\varphi$ (Explicit Factivity)

  That $\models t :^e \varphi \Rightarrow t : \varphi$ follows because $t :^e \varphi := Et \wedge t : \varphi$ (Definition 2.6). Hence $\models t :^e \varphi \Rightarrow \varphi$ by Factivity. So all that remains is to show that $\models t :^e \varphi \Rightarrow Ec_\varphi$. Proceeding, assume $w \models t :^e \varphi$. This means

  $$w \models Et \wedge t \gg \varphi \wedge \textstyle\bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta \ . \tag{16}$$

  But it follows from $w \models Et \wedge t \gg \varphi$ by Certification of Evidence (Definition 3.1) that $w \models Ec_\varphi$. Hence $\models t :^e \varphi \Rightarrow Ec_\varphi$.

- $\models B(t :^e \varphi) \Rightarrow B^e\varphi$, $\models \Box(t :^e \varphi) \Rightarrow \Box^e\varphi$, and $\models K(t :^e \varphi) \Rightarrow K^e\varphi$.

  These follow by our abbreviations (Definition 2.6) using Explicit Factivity. For example, let us show that $\models B(t :^e \varphi) \Rightarrow B^e\varphi$. Proceeding, assume $w \models B(t :^e \varphi)$. This means $w \models \Diamond\Box(t :^e \varphi)$, which itself means

  $$\exists v \leq w . \forall u \leq v : u \models t :^e \varphi \ .$$

  But it follows from $u \models t :^e \varphi$ by Explicit Factivity that $u \models Ec_\varphi$. Hence

  $$\exists v \leq w . \forall u \leq v : u \models Ec_\varphi \ .$$

  But this means $w \models \Diamond\Box Ec_\varphi$, which means $w \models B(Ec_\varphi)$, which means $w \models B^e\varphi$. Conclusion: $\models B(t :^e \varphi) \Rightarrow B^e\varphi$. Arguments for the other validities are similar.

- $\models B^e\varphi \Rightarrow B\varphi$, $\models \Box^e\varphi \Rightarrow \Box\varphi$, and $\models K^e\varphi \Rightarrow K\varphi$.

  These follow by our abbreviations using Evidential Goodness (Definition 3.7). For example, let us show that $\models B^e\varphi \Rightarrow B\varphi$. Proceeding, assume $w \models_M B^e\varphi$ for the model M. This

means $w \models_M \Diamond\Box Ec_\varphi$. The latter means that there is a $v \leq w$ such that every $u \leq v$ satisfies $u \models_M Ec_\varphi$. But from $u \models_M Ec_\varphi$ we have by Evidential Goodness (and the fact that M is a model) that $u \models_M \varphi$, and therefore we have shown that there is a $v \leq w$ such that every $u \leq v$ satisfies $u \models_M \varphi$. But the latter is what it means to have $w \models_M \Diamond\Box\varphi$, which is what it means to have $w \models_M B\varphi$. Conclusion: $\models B^e\varphi \Rightarrow B\varphi$. Arguments for the other validities are similar. □

The Internalization Property of Justification Logics [5, 10] manifests itself in various ways within the context of our framework. A proper discussion of this important property and its manifestations in our setting requires that we study other things first (including a proof system for our logic). Accordingly, we shall postpone the description and discussion of Internalization until §4.1, at which point we will be in a position to proceed.

4 Proof System

Definition 4.1 (Theory)**.** JBG, the *theory of justified belief with evidential goodness*, is defined in Table 2.

Axiom Schemes

Classical Logic:	Schemes of Classical Propositional Logic
Admissibility:	$\vdash t \gg \varphi$ whenever $t \gg \varphi$ (per Definition 2.2)
	$\vdash \neg(t \gg \varphi)$ whenever $t \not\gg \varphi$ (per Definition 2.2)
Trivial Evidence:	$\vdash Ec_\top$
Certification of Evidence:	$\vdash (Et \wedge t \gg \varphi) \Rightarrow Ec_\varphi$
Subterm Closure:	$\vdash E(t \cdot s) \vee E(t + s) \Rightarrow Et \wedge Es$
Availability of Evidence:	$\vdash E(t \cdot s) \Rightarrow K(Et \wedge Es \Rightarrow E(t \cdot s))$
	$\vdash E(t + s) \Rightarrow K(Et \wedge Es \Rightarrow E(t + s))$
Evidential Goodness:	$\vdash Ec_\varphi \Rightarrow \varphi$
Information:	S5 axioms for K
Defeasible Knowledge:	S4 axioms for $\Box$
Indefeasibility:	$\vdash K\varphi \Rightarrow \Box\varphi$
Local Connectedness:	$\vdash K(\Box\varphi \Rightarrow \psi) \vee K(\Box\psi \Rightarrow \varphi)$

Rules

$$\frac{\varphi \Rightarrow \psi \quad \varphi}{\psi}\ (\text{MP}) \qquad \frac{\varphi}{K\varphi}\ (K\text{N}) \qquad \frac{\varphi}{\Box\varphi}\ (\Box\text{N})$$

Table 2: The theory JBG

Theorem 4.2 (Soundness and Completeness, Finite Model Property)**.** $\vdash \varphi$ if and only if $\models \varphi$. Further, every satisfiable formula is satisfiable in a finite standard model.

Proof. Soundness (i.e., $\vdash \varphi$ implies $\models \varphi$) is a straightforward argument by induction on the length of derivation. We focus on the completeness argument ($\nvdash \varphi$ implies $\not\models \varphi$). We proceed by way of a canonical model argument.

Define the canonical structure $\Omega := (W^\Omega, [\![\cdot]\!]^\Omega, \sim^\Omega, \geq^\Omega, E^\Omega)$ in the usual way. That is, W^Ω is the set of all maximal JBG-consistent sets of formulas, $[\![p]\!]^\Omega := \{\Gamma \in W^\Omega \mid p \in \Gamma\}$, $\Gamma \sim^\Omega \Delta$ iff $\Gamma^K := \{\varphi \mid K\varphi \in \Gamma\} \subseteq \Delta$, $\Gamma \geq^\Omega \Delta$ iff $\Gamma^\Box := \{\varphi \mid \Box\varphi \in \Gamma\} \subseteq \Delta$, and $E^\Omega(\Gamma) := \{t \in \mathscr{T} \mid Et \in \Gamma\}$. The *Truth Lemma* (i.e., the statement that $\varphi \in \Gamma$ iff $(\Omega, \Gamma) \models \varphi$ for each formula φ) is then proved by induction on formula construction. The cases for atomic formulas, Boolean connectives, and modal formulas $K\varphi$ and $\Box\varphi$ follow by standard modal reasoning [17]; all that remains are the cases for formulas $t \gg \varphi$ and Et.

- Truth Lemma: case for formulas $t \gg \varphi$.

  We have $t \gg \varphi \in \Gamma$ iff $\vdash t \gg \varphi$ by our definition of W^Ω and the axiomatization of JBG. We have $\vdash t \gg \varphi$ iff $t \gg \varphi$ by the Admissibility axiom of JBG. We have $t \gg \varphi$ iff $(\Omega, \Gamma) \models t \gg \varphi$ by the definition of the satisfaction relation $\models$.

- Truth Lemma: case for formulas Et.

  We have $Et \in \Gamma$ iff $t \in E^\Omega(\Gamma)$ by the definition of E^Ω. We have $t \in E^\Omega(\Gamma)$ iff $(\Omega, \Gamma) \models Et$ by the definition of the satisfaction relation $\models$.

This completes the proof of the Truth Lemma. Using standard modal reasoning [17], one can then show that $\sim$ is an equivalence relation and $\geq$ is a preorder and that Ω satisfies Local Connectedness and Indefeasibility. Further, we have each of the following.

- Ω satisfies Trivial Evidence: $c_\top \in E^\Omega(\Gamma)$.

  By the Trivial Evidence axiom, the maximal JBG-consistency of Γ, and the definition of E^Ω.

- Ω satisfies Certification of Evidence: if $t \in E^\Omega(\Gamma)$ and $t \gg \varphi$, then $c_\varphi \in E^\Omega(\Gamma)$.

  Suppose $t \in E^\Omega(\Gamma)$ and $t \gg \varphi$. It follows from $t \in E^\Omega(\Gamma)$ by the definition of E^Ω that $Et \in \Gamma$. It follows from $t \gg \varphi$ by the Admissibility axiom that $t \gg \varphi \in \Gamma$. But then $Ec_\varphi \in \Gamma$ by the Certification of Evidence axiom and the maximal JBG-consistency of Γ, and hence $c_\varphi \in E^\Omega(\Gamma)$ by the definition of E^Ω.

- Ω satisfies Subterm Closure: if $\circ$ denotes either of the symbols $+$ or $\cdot$ and $t \circ s \in E^\Omega(\Gamma)$, then $t, s \in E^\Omega(\Gamma)$.

  Suppose $t \circ s \in E^\Omega(\Gamma)$. It follows that $E(t \circ s) \in \Gamma$ by the definition of E^Ω and hence that $Et, Es \in \Gamma$ by the Subterm Closure axiom and the maximal JBG-consistency of Γ. But then $t, s \in E^\Omega(\Gamma)$ by the definition of E^Ω.

- Ω satisfies Availability of Evidence: if $\circ$ denotes either of the symbols $+$ or $\cdot$, $t \circ s \in E^\Omega(\Gamma)$, $\Gamma \sim^\Omega \Gamma'$, and $t, s \in E^\Omega(\Gamma')$, then $t \circ s \in E^\Omega(\Gamma')$.

  Assume $t \circ s \in E^\Omega(\Gamma)$, $\Gamma \sim^\Omega \Gamma'$, and $t, s \in E^\Omega(\Gamma')$. Applying the definition of E^Ω, it follows from the first assumption that $E(t \circ s) \in \Gamma$ and from the third that $Et, Es \in \Gamma'$. Since $E(t \circ s) \in \Gamma$, it follows by the maximal JBG-consistency of Γ and the Availability of Evidence axiom that $K(Et \wedge Es \Rightarrow E(t \circ s)) \in \Gamma$. Since $\Gamma \sim^\Omega \Gamma'$, it follows that $Et \wedge Es \Rightarrow E(t \circ s) \in \Gamma'$. But it follows from $Et \wedge Es \Rightarrow E(t \circ s) \in \Gamma'$ and $Et, Es \in \Gamma'$ by the maximal JBG-consistency of Γ' that $E(t \circ s) \in \Gamma'$, and the latter implies $t \circ s \in E^\Omega(\Gamma')$ by the definition of E^Ω.

- Ω satisfies Evidential Goodness: $c_\varphi \in E^\Omega(\Gamma)$ implies $\Gamma \models \varphi$.

  Suppose $c_\varphi \in E^\Omega(\Gamma)$. Applying the definition of E^Ω, we have that $Ec_\varphi \in \Gamma$ and hence that $\varphi \in \Gamma$ by the Evidential Goodness axiom and the maximal JBG-consistency of Γ. But it follows from $\varphi \in \Gamma$ by the Truth Lemma that $\Gamma \models \varphi$.

We have shown that Ω is a model.

We now fix a formula φ for which $\nvdash \varphi$. Clearly, we may extend $\{\neg\varphi\}$ by way of a Lindenbaum argument [17] to a maximal JBG-consistent set $\Gamma^* \in W^\Omega$ and then we have $(\Omega, \Gamma^*) \not\models \varphi$ by the Truth Lemma. Unfortunately, this does not show that $\not\models \varphi$ because the canonical structure Ω, while a model, need not be a *standard* model. So it will now be our task to produce a standard pointed model at which φ does not hold. This will then show that $\not\models \varphi$ and thereby complete the proof.

Define Ω' to be the submodel of Ω obtained removing all worlds not $\sim^\Omega$-connected to Γ^* and restricting the other components of Ω to the remaining worlds. That is, $\Omega' := (W^{\Omega'}, [\![\cdot]\!]^{\Omega'}, \sim^{\Omega'}, \geq^{\Omega'}, E^{\Omega'})$, where $W^{\Omega'} := \{\Gamma \in W^\Omega \mid \Gamma \sim^\Omega \Gamma^*\}$, $[\![p]\!]^{\Omega'} := [\![p]\!]^\Omega \cap W^{\Omega'}$, $\sim^{\Omega'} := \sim^\Omega \cap (W^{\Omega'} \times W^{\Omega'})$, $\geq^{\Omega'} := \geq^\Omega \cap (W^{\Omega'} \times W^{\Omega'})$, and $E^{\Omega'}(\Gamma) := E^\Omega(\Gamma)$ for each $\Gamma \in W^{\Omega'}$. Note that $\sim^{\Omega'}$ is *universal*; that is, $\sim^{\Omega'} = W^{\Omega'} \times W^{\Omega'}$. Furthermore, transitioning from Ω to Ω' preserves satisfaction: for each formula ψ and world $\Gamma \in W^{\Omega'}$, we have $(\Omega', \Gamma) \models \psi$ iff $(\Omega, \Gamma) \models \psi$. It follows that Ω' is a model (though not necessarily a *standard* model) and $(\Omega', \Gamma^*) \not\models \varphi$.

For each formula $\theta \in \mathscr{F}$, define the set $\mathsf{rsub}(\theta)$ of *recursive subformulas* of θ according to the following induction: $\mathsf{rsub}(\top) := \{\top\}$, $\mathsf{rsub}(p) := \{p\}$ for each $p \in \Phi$, $\mathsf{rsub}(\neg\theta) := \{\neg\theta\} \cup \mathsf{rsub}(\theta)$, $\mathsf{rsub}(\theta_1 \wedge \theta_2) := \{\theta_1 \wedge \theta_2\} \cup \mathsf{rsub}(\theta_1) \cup \mathsf{rsub}(\theta_2)$, $\mathsf{rsub}(K\theta) := \{K\theta\} \cup \mathsf{rsub}(\theta)$, $\mathsf{rsub}(\Box\theta) := \{\Box\theta\} \cup \mathsf{rsub}(\theta)$,

$$\begin{aligned} \mathsf{rsub}(t \gg \theta) &:= \{t \gg \theta\} \cup \mathsf{rsub}(\theta) \cup \bigcup_{c_\chi \in \mathsf{sub}(t)} \mathsf{rsub}(\chi) \text{ , and} \\ \mathsf{rsub}(Et) &:= \{Et\} \cup \bigcup_{c_\chi \in \mathsf{sub}(t)} \mathsf{rsub}(\chi) \text{ .} \end{aligned}$$

Define the language $\mathscr{L}_0^\varphi := (\mathscr{T}_0^\varphi, \mathscr{F}_0^\varphi)$ by

$$\begin{aligned} \mathscr{T}_0^\varphi &:= \{c_\top\} \\ \mathscr{F}_0^\varphi &:= \{\top, Ec_\top, c_\top \gg \top\} \cup (\Phi \cap \mathsf{rsub}(\varphi)) \end{aligned}$$

Then, whenever $\mathscr{L}_i^\varphi$ is defined, define the language $\mathscr{L}_{i+1}^\varphi := (\mathscr{T}_{i+1}^\varphi, \mathscr{F}_{i+1}^\varphi)$ consisting of the set $\mathscr{T}_{i+1}^\varphi$ of the terms t_{i+1} and the set $\mathscr{F}_{i+1}^\varphi$ of formulas θ_{i+1} defined by the following grammar:

$$\begin{aligned} \theta_{i+1} &::= \theta_i \mid \neg\theta_i \mid \theta_i \wedge \theta_i \mid K\theta_i \mid \Box\theta_i \mid Et_{i+1} \mid t_{i+1} \gg \theta_i \\ t_{i+1} &::= t_i \mid c_{\theta_i} \mid t_i \cdot t_i \mid t_i + t_i \\ &\quad \text{where } \theta_i \in \mathscr{F}_i^\varphi \text{ and } t_i \in \mathscr{T}_i^\varphi \end{aligned}$$

One can show by induction on $i \in \mathbb{N}$ that $\mathscr{F}_i^\varphi$ and $\mathscr{T}_i^\varphi$ are both finite, that $\mathscr{F}_i^\varphi$ is closed under subformulas (i.e., $\psi \in \mathscr{F}_i^\varphi$ implies $\mathsf{sub}(\psi) \subseteq \mathscr{F}_i^\varphi$, which is argued by a sub-induction on the construction of ψ), that $\mathscr{T}_i^\varphi$ is closed under subterms (i.e., $t \in \mathscr{T}_i^\varphi$ implies $\mathsf{sub}(t) \subseteq \mathscr{T}_i^\varphi$, which is argued by a sub-induction on the construction of t), and that $\bigcup_{j=0}^{i-1} \mathscr{F}_j^\varphi \subseteq \mathscr{F}_i^\varphi$ and $\bigcup_{j=0}^{i-1} \mathscr{T}_j^\varphi \subseteq \mathscr{T}_i^\varphi$.

Let m be the smallest $i \in \mathbb{N}$ such that $\varphi \in \mathscr{F}_i^\varphi$. (That such an m exists is shown by induction on the construction of φ.) Define the language $\mathscr{L}^* := (\mathscr{T}^*, \mathscr{F}^*) := (\mathscr{T}_m^\varphi, \mathscr{F}_m^\varphi)$. Define the equivalence

relation $\equiv$ on $W^{\Omega'}$ by $\Gamma \equiv \Delta$ iff $\Gamma \cap \mathscr{F}^* = \Delta \cap \mathscr{F}^*$ and let $[\Gamma] := \{\Delta \in W^{\Omega'} \mid \Delta \equiv \Gamma\}$ be the equivalence class of Γ under this relation. Define $M := (W, [\![\cdot]\!], \sim, \geq, E)$ by

$$\begin{aligned} W &:= \{[\Gamma] \mid \Gamma \in W^{\Omega'}\} \ , \\ [\![p]\!] &:= \{[\Gamma] \in W \mid p \in \Phi \cap \mathscr{F}^* \text{ and } \Gamma \in [\![p]\!]^{\Omega'}\} \ , \end{aligned}$$

$\sim$ is "smallest filtration" given by

$$[\Gamma] \sim [\Delta] \quad \text{iff} \quad \text{there exists } (\Gamma', \Delta') \in [\Gamma] \times [\Delta] \text{ such that } \Gamma' \sim^{\Omega'} \Delta' \ ,$$

$\geq$ is the transitive filtration given by

$$[\Gamma] \geq [\Delta] \quad \text{iff} \quad \forall \Box\psi \in \mathscr{F}^*(\Box\psi \in \Gamma \Rightarrow \Box\psi \in \Delta) \ ,$$

and $E([\Gamma]) := \mathscr{T}^* \cap \bigcap_{\Gamma' \in [\Gamma]} E^{\Omega'}(\Gamma')$. We observe that since $\sim^{\Omega'}$ is universal, so is $\sim$; that is, $\sim = W \times W$. Furthermore, M is finite: the set W of worlds is finite (since there is a bijection mapping each world $[\Gamma]$ to a subset $\Gamma \cap \mathscr{F}^*$ of the finite set $\mathscr{F}^*$), the available evidence $E([\Gamma])$ at each world $[\Gamma] \in W$ is finite (since $E([\Gamma]) \subseteq \mathscr{T}^*$ and $\mathscr{T}^*$ is finite), and there are at most finitely many letters $p \in \Phi$ that are true at any given world $[\Gamma] \in W$ (since $[\Gamma] \in [\![p]\!]$ implies $p \in \mathscr{F}^*$ and $\mathscr{F}^*$ is finite).

We wish to show that E satisfies the more convenient equality

$$E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma) \ .$$

To prove this, we need to show that $\Gamma_1 \equiv \Gamma_2$ implies $\mathscr{T}^* \cap E^{\Omega'}(\Gamma_1) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma_2)$ because then we have

$$E([\Gamma]) = \mathscr{T}^* \cap \bigcap_{\Gamma' \in [\Gamma]} E^{\Omega'}(\Gamma') = \bigcap_{\Gamma' \in [\Gamma]} \mathscr{T}^* \cap E^{\Omega'}(\Gamma') = \mathscr{T}^* \cap E^{\Omega'}(\Gamma) \ .$$

Proceeding, assume that $\Gamma_1 \equiv \Gamma_2$ and $t \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma_1)$. Since $t \in \mathscr{T}^* = \mathscr{T}^{\varphi}_m$, it follows by the grammar defining $\mathscr{L}^{\varphi}_m$ that $Et \in \mathscr{F}^{\varphi}_m = \mathscr{F}^*$. Since $t \in E^{\Omega'}(\Gamma_1)$, it follows by the definition of $E^{\Omega'}(\Gamma_1)$ (in terms of $E^{\Omega}(\Gamma_1)$) that $Et \in \Gamma_1$. Hence $Et \in \mathscr{F}^* \cap \Gamma_1$. But $\Gamma_1 \equiv \Gamma_2$ means that $\Gamma_1 \cap \mathscr{F}^* = \Gamma_2 \cap \mathscr{F}^*$, and therefore $Et \in \mathscr{F}^* \cap \Gamma_2$. It follows from $Et \in \mathscr{F}^* \cap \Gamma_2$ that $Et \in \Gamma_2$ and therefore that $t \in E^{\Omega'}(\Gamma_2)$ by the definition of $E^{\Omega'}(\Gamma_2)$ (in terms of $E^{\Omega}(\Gamma_2)$). Hence $t \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma_2)$. The argument that $\Gamma_1 \equiv \Gamma_2$ and $t \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma_2)$ together imply that $t \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma_1)$ is proved similarly. Conclusion: $E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$.

We show that M satisfies Trivial Evidence. Choose a $[\Gamma] \in W$ (we wish to show that $c_\top \in E([\Gamma])$). Since Ω' satisfies Trivial Evidence, we have $c_\top \in E^{\Omega'}(\Gamma)$. Further, since $c_\top \in \mathscr{T}^{\varphi}_0$, we have $c_\top \in \mathscr{T}^{\varphi}_m = \mathscr{T}^*$. Hence $c_\top \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma) = E([\Gamma])$. Conclusion: M satisfies Trivial Evidence.

We show that for each $i \in \mathbb{N}$, if $t \in \mathscr{T}^{\varphi}_i$ and $t \gg \psi$, then $c_\psi \in \mathscr{T}^{\varphi}_i$. This is proved by an induction on $i \in \mathbb{N}$. The case $i = 0$ is verified immediately (we have that $\mathscr{T}^{\varphi}_0 = \{c_\top\}$ and that $c_\top \gg \chi$ implies $\chi = \top$). So we assume the result holds for all $k < i$ and prove it holds for $k = i > 0$ by a sub-induction on the construction of t. For the sub-induction base, we assume $c_\chi \in \mathscr{T}^{\varphi}_i$ and $c_\chi \gg \psi$. It follows by the definition of admissibility (Definition 2.2) that $\psi = \chi$. But $c_\chi \in \mathscr{T}^{\varphi}_i$ implies $c_\chi \in \mathscr{T}^{\varphi}_i$ (trivially). This completes the sub-induction base. For the sub-induction inductive step, we consider the cases $u \cdot v \in \mathscr{T}^{\varphi}_i$ and $u + v \in \mathscr{T}^{\varphi}_i$ separately. First, we assume $u \cdot v \in \mathscr{T}^{\varphi}_i$ and $(u \cdot v) \gg \psi$. It follows by the definition of admissibility that there exists a formula χ such that $u \gg (\chi \Rightarrow \psi)$ and $v \gg \chi$. Further, it follows by the grammar for $\mathscr{L}^{\varphi}_i$ that $u \cdot v \in \mathscr{T}^{\varphi}_i$ implies $u \in \mathscr{T}^{\varphi}_i$. By the sub-induction hypothesis for u, we have $c_{\chi \Rightarrow \psi} \in \mathscr{T}^{\varphi}_i$. But $c_{\chi \Rightarrow \psi} \in \mathscr{T}^{\varphi}_i$ implies $\chi \Rightarrow \psi \in \mathscr{F}^{\varphi}_{i-1}$

(by the grammar for $\mathscr{L}_i^\varphi$), and the latter implies $\psi \in \mathscr{F}_{i-1}^\varphi$ (by the fact that $\mathscr{F}_{i-1}^\varphi$ is closed under subformulas). It follows by the grammar for $\mathscr{L}_i^\varphi$ that $c_\psi \in \mathscr{T}_i^\varphi$, which completes the argument for this case. We now consider the final case by assuming that $u+v \in \mathscr{T}_i^\varphi$ and $(u+v) \gg \psi$. It follows by the definition of admissibility that there is a formula χ such that, without loss of generality, $u \gg \psi$ and $v \gg \chi$. Further, it follows by the grammar for $\mathscr{L}_i^\varphi$ that $u + v \in \mathscr{T}_i^\varphi$ implies $u \in \mathscr{T}_i^\varphi$. By the sub-induction hypothesis for u, we have $c_\psi \in \mathscr{T}_i^\varphi$, as desired. This completes the sub-induction and the overall induction. A particular consequence of the argument of this paragraph is the following: if $t \in \mathscr{T}^*$ and $t \gg \psi$, then $c_\psi \in \mathscr{T}^*$.

We now show that M satisfies Certification of Evidence. Proceeding, we assume $t \in E([\Gamma])$ and $t \gg \psi$ (and we wish to show that $c_\psi \in E([\Gamma])$). Since $t \in E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$, it follows from the assumption $t \gg \psi$ and the result at the end of the previous paragraph that $c_\psi \in \mathscr{T}^*$. Also, since Ω' satisfies Certification of Evidence, $t \in E^{\Omega'}(\Gamma)$ and $t \gg \psi$ together imply that $c_\psi \in E^{\Omega'}(\Gamma)$. Hence $c_\psi \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma) = E([\Gamma])$. So M satisfies Certification of Evidence.

We now show that M satisfies Subterm Closure. Proceeding, we take $\circ$ to be either of the symbols $+$ or $\cdot$ and assume that $t \circ s \in E([\Gamma])$ (with our goal being to show that $t, s \in E([\Gamma])$). But $t \circ s \in E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$, the set $\mathscr{T}^*$ is closed under subterms, and Ω' satisfies Subterm Closure (so that $t, s \in E^{\Omega'}([\Gamma])$). The result follows.

We now show that M satisfies Availability of Evidence. Proceeding, we take $\circ$ to be either of the symbols $+$ or $\cdot$ and assume $t \circ s \in E([\Gamma])$, $[\Gamma] \sim [\Delta]$, and $t, s \in E([\Delta])$ (with our goal being to show that $t \circ s \in E([\Delta])$). Now we have $t \circ s \in E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$ and $t, s \in E([\Delta]) = \mathscr{T}^* \cap E^{\Omega'}(\Delta)$. Since $\sim^{\Omega'}$ is universal, we also have $\Gamma \sim^{\Omega'} \Delta$. But Ω' satisfies Availability of Evidence and therefore $t \circ s \in E^{\Omega'}(\Delta)$. Hence $t \circ s \in \mathscr{T}^* \cap E^{\Omega'}(\Delta) = E([\Delta])$, which is what we wished to show.

$\sim$ is an equivalence relation because it is universal, and $\geq$ is a preorder because we took the transitive filtration of a reflexive relation and $\Box$ is $\mathsf{S4}$ [22, pp. 105–107]. Indefeasibility follows by the universality of $\sim$.

To see that M satisfies Local Connectedness (i.e., that $[\Gamma] \sim [\Delta]$ implies $[\Gamma] \geq [\Delta]$ or $[\Delta] \geq [\Gamma]$), assume toward a contradiction that $[\Gamma] \sim [\Delta]$, $[\Gamma] \not\geq [\Delta]$, and $[\Delta] \not\geq [\Gamma]$. It follows by the meaning of $\not\geq$ (per the negation of the definition of $\geq$) that there are formulas $\Box\psi, \Box\psi' \in \mathscr{F}^*$ such that $\Box\psi \in \Gamma$, $\Box\psi \notin \Delta$, $\Box\psi' \in \Delta$, and $\Box\psi' \notin \Gamma$. By the Truth Lemma (and the preservation of the satisfaction relation in going from Ω to Ω'), we have $(\Omega', \Gamma) \models \Box\psi \wedge \neg\Box\psi'$ and $(\Omega', \Delta) \models \Box\psi' \wedge \neg\Box\psi$. Since $\geq^{\Omega'}$ is transitive, it follows that $(\Omega', \Gamma) \models \Box\Box\psi \wedge \neg\Box\psi'$ and $(\Omega', \Delta) \models \Box\Box\psi' \wedge \neg\Box\psi$. There therefore exist worlds $\Gamma' \leq^{\Omega'} \Gamma$ and $\Delta' \leq^{\Omega'} \Delta$ such that $(\Omega', \Gamma') \models \Box\psi \wedge \neg\psi'$ and $(\Omega', \Delta') \models \Box\psi' \wedge \neg\psi$. Since $\sim^{\Omega'}$ is universal, we have $\Gamma' \sim^{\Omega'} \Gamma'$ and $\Gamma' \sim^{\Omega'} \Delta'$. But then $(\Omega', \Gamma') \not\models K(\Box\psi \Rightarrow \psi') \vee K(\Box\psi' \Rightarrow \psi)$, from which it follows by the Truth Lemma (and the preservation of satisfaction in going from Ω to Ω') that $K(\Box\psi \Rightarrow \psi') \vee K(\Box\psi' \Rightarrow \psi) \notin \Gamma'$. But the latter contradicts the maximal consistency of Γ' because $K(\Box\psi \Rightarrow \psi') \vee K(\Box\psi' \Rightarrow \psi)$ is the Local Connectedness axiom. Conclusion: our original assumption (that M does not satisfy Local Connectedness) was incorrect, and hence M does in fact satisfy Local Connectedness.

We have proved that M is a pre-model. Further, it follows immediately from the finiteness of M that M satisfies standardness. So M is a standard pre-model.

We now prove the *Filtration Lemma*: for each $[\Gamma] \in W$ and each $\psi \in \mathscr{F}^*$, we have $(M, [\Gamma]) \models \psi$ iff $(\Omega', \Gamma) \models \psi$. The argument is by induction on the construction of $\psi \in \mathscr{F}^*$. The cases for atomic formulas, Boolean connectives, and modal formulas $K\varphi$ and $\Box\varphi$ follow by standard filtration arguments [22, pp. 101–102]. All that remains are the cases for formulas $t \gg \psi \in \mathscr{F}^*$ and $Et \in \mathscr{F}^*$.

- Filtration Lemma: case for formulas $t \gg \psi \in \mathscr{F}^*$.

$(M, [\Gamma]) \models t \gg \psi$ means $t \gg \psi$, which is what it means to have $(\Omega', \Gamma) \models t \gg \psi$.

- Filtration Lemma: case for formulas $Et \in \mathscr{F}^*$.

  Assume $(M, [\Gamma]) \models Et$. It follows that $t \in E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$. But $t \in E^{\Omega'}(\Gamma)$ is what it means to have $(\Omega', \Gamma) \models Et$.

  Conversely, assume $(\Omega', \Gamma) \models Et$, which means that $t \in E^{\Omega'}(\Gamma)$. Since this case assumes $Et \in \mathscr{F}^* = \mathscr{F}^{\varphi}_m$, it follows by the grammar for $\mathscr{L}^{\varphi}_m$ that $t \in \mathscr{T}^{\varphi}_m = \mathscr{T}^*$ (check this both for $\mathscr{L}^{\varphi}_0$ and for $\mathscr{L}^{\varphi}_i$ with $i > 0$). But then $t \in \mathscr{T}^* \cap E^{\Omega'}(\Gamma) = E([\Gamma])$, which is what it means to have $(M, [\Gamma]) \models Et$.

This completes the proof of the Filtration Lemma.

We now prove that M satisfies Evidential Goodness. Proceeding, we assume that $c_\psi \in E([\Gamma])$ (and wish to show that $(M, [\Gamma]) \models \psi$). Now $c_\psi \in E([\Gamma]) = \mathscr{T}^* \cap E^{\Omega'}(\Gamma)$ implies $c_\psi \in \mathscr{T}^* = \mathscr{T}^{\varphi}_m$, from which it follows by the grammar for $\mathscr{L}^{\varphi}_m$ that $\psi \in \mathscr{F}^{\varphi}_m = \mathscr{F}^*$ (check this both for $\mathscr{L}^{\varphi}_0$ and for $\mathscr{L}^{\varphi}_i$ with $i > 0$). Further, $c_\psi \in E^{\Omega'}(\Gamma)$ implies $(\Omega', \Gamma) \models \psi$ by the fact that Ω' satisfies Evidential Goodness. But from $\psi \in \mathscr{F}^*$ and $(\Omega', \Gamma) \models \psi$, it follows by the Filtration Lemma that $(M, [\Gamma]) \models \psi$, which is what we wished to show. Conclusion: M satisfies Evidential Goodness.

A standard pre-model satisfying Evidential Goodness is a standard model. So we have shown that M is a standard model. Further, it follows from $(\Omega', \Gamma^*) \not\models \varphi$ and $\varphi \in \mathscr{F}^*$ by the Filtration Lemma that $(M, [\Gamma^*]) \not\models \varphi$. □

Corollary 4.3. The satisfiability problem for JBG is decidable.

4.1 Internalization

The Internalization Property of a Justification Logic states that for every provable formula φ, there exists a term t such that $t : \varphi$ is also provable [10]. Intuitively, this property says that the Justification Logic is "aware of" (or "internalizes") its own proofs, by which we mean that any proof of φ may be represented as a term t to which the logic accords status as justification for φ (i.e., $t : \varphi$ is also provable).

Our logic also satisfies Internalization, though our broader language allows us to state a number of different kinds of Internalization. We state these in Theorems 4.7 and 4.9 after considering a few preliminary matters.

Definition 4.4 (Necessitated Axioms and Logical Terms)**.** A *necessitated axiom* is a formula of the form

$$\underbrace{X_1 X_2 X_3 \cdots X_n}_{\text{zero or more } X_i\text{'s}} \varphi \tag{17}$$

where each $X_i \in \{\Box, K\}$ and φ is a JBG-axiom. A *necessitated non-axiom* is a formula of the form (17) for which n is maximal (i.e., φ is not itself of the form $X\psi$ for some $X \in \{\Box, K\}$ and formula ψ) and φ is not a JBG-axiom. The set of *logical terms* is the smallest set that contains certificates c_φ for each necessitated axiom φ and is closed under the evidence-combining operation $t, s \mapsto t \cdot s$.

Definition 4.5 (JBG-proofs, Necessitations, $\pi \vdash_n \varphi$, $\vdash^* \varphi$)**.** A JBG-*proof* is a finite nonempty sequence π of formulas such that every formula in π is either an instance of a JBG-axiom scheme or else follows from formulas occurring earlier in the sequence by a JBG-rule. A *line* of a JBG-proof is a formula occurring in that proof. To say a JBG-proof π is "of" a formula φ means that φ is the

last line of π. The *length* of a JBG-proof π, denoted $|\pi|$, is the number of lines in π. If π and π' are JBG-proofs, then $\pi \supseteq \pi'$ means that π contains every line in π'. A *necessitation* in a JBG-proof π is a non-axiom φ occurring in π that can only have followed from an earlier formula by KN (the "necessitation rule for K") or by $\Box$N (the "necessitation rule for $\Box$").[8] Given a JBG-proof π, a *necessitation of a necessitated non-axiom* is a necessitation φ that has the form of a necessitated non-axiom (Definition 4.4). For a non-negative integer n, we write $\pi \vdash_n \varphi$ to mean that π is a JBG-proof of φ that contains at most n necessitations of a necessitated non-axiom. Finally, we write $\vdash^* \varphi$ to mean that there exists a π such that $\pi \vdash_0 \varphi$.

Lemma 4.6 (Necessitation Elimination)**.** For each $\varphi \in \mathscr{F}$, we have $\vdash \varphi$ iff $\vdash^* \varphi$.

Proof. $\vdash^* \varphi$ implies $\vdash \varphi$ trivially. So all that remains is to prove the converse. To prove this, it suffices for us to prove that $\pi \vdash_n \varphi$ implies there exists $\pi_* \supseteq \pi$ such that $\pi_* \vdash_0 \varphi$. We prove this by induction on n with a sub-induction on the length $|\pi|$ of π. Within this argument, a "proof" is a JBG-proof.

- *Base case:* $n = 0$. The result follows immediately (take $\pi_* = \pi$).

- *Induction step:* assume the result holds for up to $n-1$ necessitations of a necessitated non-axiom (the "induction hypothesis"), prove it holds for up to n necessitations of a necessitated non-axiom. We proceed by a sub-induction on the length $|\pi|$ of a proof π.

  Sub-induction base: $|\pi| = 1$. Assume $\pi \vdash_n \varphi$. Since π contains only one line, φ is an axiom. Hence $\pi \vdash_0 \varphi$. Taking $\pi_* = \pi$, we have $\pi_* \supseteq \pi$ with $\pi_* \vdash_0 \varphi$.

  Sub-induction step: assume the result holds for proofs having fewer than $|\pi|$ lines (the "sub-induction hypothesis"), prove it holds for the proof π made up of exactly $|\pi|$ lines. Proceeding, we assume that $\pi \vdash_n \varphi$. If it happens that we also have $\pi \vdash_0 \varphi$, then the result follows immediately (take $\pi_* = \pi$). So let us assume further that $\pi \nvdash_0 \varphi$. That is, the proof

  $$\pi = \theta_1, \ldots, \theta_{|\pi|-1}, \varphi$$

  uses at least one necessitation of a necessitated non-axiom. Using the symbol X to denote one of K or $\Box$, let π' be the shortest prefix

  $$\pi' = \theta_1, \ldots, \theta_{|\pi'|-1}, X\theta_m$$

  of π whose last line $X\theta_m$ is a necessitation of a necessitated non-axiom. That such a nonempty π' exists follows because π contains at least one necessitation of a necessitated non-axiom. So either π' is a proper prefix of π or $\pi' = \pi$. Let us consider each case in turn.

  Case: π' is a proper prefix of π. Hence $|\pi'| < |\pi|$ and $\pi = \pi'.\sigma$ (with "." denoting sequence concatenation) for some finite nonempty sequence σ of formulas, and so we may apply the sub-induction hypothesis: there exists a $\pi'_* \supseteq \pi'$ such that $\pi'_* \vdash_0 X\theta_m$. That is, π'_* is a proof of $X\theta_m$ that contains every line in π' and that contains no necessitations of a necessitated non-axiom, and therefore $\pi'_*.\sigma$ is a proof of φ that contains every line in $\pi = \pi'.\sigma$ and that

[8] So φ is *not* a necessitation in a JBG-proof π whenever it is possible to derive φ from earlier lines in π using MP. In particular, if φ can be derived from earlier lines in π both by MP and by one of the necessitation rules, then φ is *not* a necessitation.

contains at most $n-1$ necessitations of a necessitated non-axiom.[9] That is, $\pi'_*.\sigma \vdash_{n-1} \varphi$ with $\pi'_*.\sigma \supseteq \pi'.\sigma = \pi$. Applying the induction hypothesis, there exists a proof $\pi_* \supseteq \pi'_*.\sigma \supseteq \pi'.\sigma = \pi$ such that $\pi_* \vdash_0 \varphi$, as desired.

Case: $\pi' = \pi$. Hence $\varphi = X\theta_m$. By the definition of π' (as the *shortest* prefix of π whose last line is a necessitation of a necessitated non-axiom) and the fact that $\pi' = \pi$, it follows that $\pi \vdash_1 X\theta_m$. Therefore, θ_m must itself be a necessitated non-axiom that is derived by MP from formulas $\theta_n \Rightarrow \theta_m$ and θ_n that appear in π before line m. That is, π has the form

$$\begin{array}{rcll} \pi & = & \sigma\ ,\ \theta_m\ ,\ \tau\ ,\ X\theta_m & \\ & & \text{with both } \theta_n \Rightarrow \theta_m \text{ and } \theta_n \text{ in } \sigma & \end{array}$$

where σ and τ are placeholders for finite (possibly empty) sequences of formulas. It follows by the necessitation rule for X and the fact that $\theta_n \Rightarrow \theta_m$ and θ_n are both in σ that we have proofs

$$\begin{array}{rcll} \pi''_1 & = & \sigma\ ,\ X\theta_n & \text{with } |\pi''_1| \leq m \leq |\pi| - 1 \\ \pi''_2 & = & \sigma\ ,\ X(\theta_n \Rightarrow \theta_m) & \text{with } |\pi''_2| \leq m \leq |\pi| - 1 \end{array}$$

that have at most one necessitation of a necessitated non-axiom. That is, $\pi''_1 \vdash_1 X\theta_n$ and $\pi''_2 \vdash_1 X(\theta_n \Rightarrow \theta_m)$. Since $|\pi''_1| < |\pi|$ and $|\pi''_2| < |\pi|$, it follows by the sub-induction hypothesis that there exists $\pi^1_* \supseteq \pi''_1$ and $\pi^2_* \supseteq \pi''_2$ such that $\pi^1_* \vdash_0 X\theta_n$ and $\pi^2_* \vdash_0 X(\theta_n \Rightarrow \theta_m)$. Making use of the Kripke axiom $X(\theta_n \Rightarrow \theta_m) \Rightarrow (X\theta_n \Rightarrow X\theta_m)$, which we have for each $X \in \{\Box, K\}$ because $\Box$ is $\mathsf{S4}$ and K is $\mathsf{S5}$, the sequence

$$\begin{array}{rl} \pi_* = & \pi^1_*\ ,\ \pi^2_*\ ,\ \theta_m\ ,\ \tau\ ,\ X(\theta_n \Rightarrow \theta_m) \Rightarrow (X\theta_n \Rightarrow X\theta_m)\ , \\ & X\theta_n \Rightarrow X\theta_m\ ,\ X\theta_m \end{array}$$

is a proof of $X\theta_m = \varphi$ that contains no necessitations of a necessitated non-axiom and that contains all formulas in π.[10] That is, $\pi_* \vdash_0 \varphi$ with $\pi_* \supseteq \pi$, as desired. □

Theorem 4.7 (Internalization)**.** For each $\varphi \in \mathscr{F}$:

$$\vdash \varphi \quad \text{iff} \quad \vdash t : \varphi \text{ for some logical } t\ .$$

Proof. We first prove the left-to-right direction. By Lemma 4.6, it suffices to prove by induction on derivation length that $\vdash^* \varphi$ implies $\vdash^* t : \varphi$ for some logical t.

- Case: φ is a necessitated axiom.

  Since φ is a necessitated axiom, we have that c_φ is logical. From $\vdash^* \varphi$, it follows by $\mathsf{sub}(c_\varphi) = \{c_\varphi\}$ that $\vdash^* \bigwedge_{c_\theta \in \mathsf{sub}(c_\varphi)} \theta$. But then since $\vdash^* c_\varphi \gg \varphi$, we have shown that $\vdash^* c_\varphi : \varphi$.

[9] Note that $\pi'_*.\sigma$ is indeed a proof: any formula in σ derived by a rule that made use of one or more lines in π' can make use of the same lines in π'_* because $\pi'_* \supseteq \pi'$.

[10] Note that π_* is indeed a proof: π^1_* and π^2_* are proofs; σ contains both $\theta_n \Rightarrow \theta_m$ and θ_n and so θ_m follows by MP from $\theta_n \Rightarrow \theta_m$ and θ_n in $\pi^1_* \supseteq \sigma$; any formula in τ that follows in π by a rule from lines in $\sigma \cup \{\theta_m\}$ follows in π_* by the same rule from the same lines in $\pi^2_* \cup \{\theta_m\} \supseteq \sigma \cup \{\theta_m\}$; the formula $X(\theta_n \Rightarrow \theta_m) \Rightarrow (X\theta_n \Rightarrow X\theta_m)$ is an axiom; the formula $X\theta_n \Rightarrow X\theta_m$ follows by MP from the line that precedes it in π_* and the last line of the proof π^2_*; and the formula $X\theta_m$ follows by MP from the line that precedes it in π_* and the last line of the proof π^1_*.

- Case: φ follows by MP from $\psi \Rightarrow \varphi$ and ψ.

  By the induction hypothesis, there are logical t and s such that $\vdash^* t : (\psi \Rightarrow \varphi)$ and $\vdash^* s : \psi$. This means

$$\vdash^* t \gg (\psi \Rightarrow \varphi) \wedge \bigwedge_{c_\theta \in \mathsf{sub}(t)} \theta \ , \tag{18}$$

$$\vdash^* s \gg \psi \wedge \bigwedge_{c_\theta \in \mathsf{sub}(s)} \theta \ . \tag{19}$$

  It follows that $\vdash^* (t \cdot s) \gg \varphi$. Further, since $\mathsf{sub}(t \cdot s) = \{t \cdot s\} \cup \mathsf{sub}(t) \cup \mathsf{sub}(s)$, it follows by (18) and (19) that $\vdash^* \bigwedge_{c_\theta \in \mathsf{sub}(t \cdot s)} \theta$. But then we have shown that $\vdash^* (t \cdot s) : \varphi$.

This completes the left-to-right direction. The right-to-left direction follows by soundness, the fact that $\models t : \varphi \Rightarrow \varphi$ (Proposition 3.12), and completeness. □

Lemma 4.8 (Certification)**.** For each pointed pre-model (M, w), formula $\varphi \in \mathscr{F}$, and propositional attitude symbol $X \in \{B, \Box, K\}$:

$$w \models X(c_\varphi : \varphi) \quad \text{iff} \quad w \models X(t : \varphi) \text{ for some } t$$
$$w \models X(c_\varphi :^e \varphi) \quad \text{iff} \quad w \models X(t :^e \varphi) \text{ for some } t$$

Proof. The left-to-right directions are obvious, so we only address the right-to-left directions. For implicit belief ($X = B$), we assume $w \models B(t : \varphi)$. This means

$$\exists v \leq w . \forall u \leq v : u \models t : \varphi \ . \tag{20}$$

We have $\models t : \varphi \Rightarrow \varphi$ by Factivity (Proposition 3.12); also, we have $c_\varphi \gg \varphi$ (by Definition 2.2) and $\bigwedge_{c_\theta \in \mathsf{sub}(c_\varphi)} \theta = \varphi$ (because $\mathsf{sub}(c_\varphi) = \{c_\varphi\}$). Therefore, applying (20) and the meaning of $c_\varphi : \varphi$ (Definition 2.6), it follows that

$$\exists v \leq w . \forall u \leq v : u \models c_\varphi : \varphi \ . \tag{21}$$

But this means $w \models B(c_\varphi : \varphi)$. For explicit belief ($X = B^e$), we assume $w \models B(t :^e \varphi)$. This means

$$\exists v \leq w . \forall u \leq v : u \models t :^e \varphi \ , \tag{22}$$

which itself means

$$\exists v \leq w . \forall u \leq v : u \models Et \wedge t : \varphi \ . \tag{23}$$

Since $\models t : \varphi \Rightarrow t \gg \varphi$ (Definition 2.6), it follows from $u \models Et$ by Certification of Evidence (Definition 3.1) and the definition of satisfaction that $u \models Ec_\varphi$. Also, it follows by our argument for the implicit case that $u \models t : \varphi$ implies $u \models c_\varphi : \varphi$. Therefore, we have shown that

$$\exists v \leq w . \forall u \leq v : u \models Ec_\varphi \wedge c_\varphi : \varphi \ , \tag{24}$$

which is what it means to have $w \models B(c_\varphi :^e \varphi)$.

For $X = \Box$, change the quantifiers "$\exists v \leq w . \forall u \leq v$" in the argument for implicit belief to "$\forall v \leq w$" and the B's to $\Box$'s. For $X = K$, change the quantifiers in the argument for implicit belief to "$\forall v \sim w$" and the B's to K's. □

Theorem 4.9 (Knowledge and Belief Internalization)**.** For each pointed model (M, w), formula $\varphi \in \mathscr{F}$, and propositional attitude symbol $X \in \{B, \Box, K\}$:

$$\begin{aligned} w \models X\varphi \quad &\text{iff} \quad w \models X(t:\varphi) \text{ for some } t \\ w \models X^e\varphi \quad &\text{iff} \quad w \models X(t:^e \varphi) \text{ for some } t \end{aligned}$$

Proof. For implicit belief, $w \models B\varphi$ means

$$\exists v \leq w. \forall u \leq v : u \models \varphi \ . \tag{25}$$

Since $c_\varphi \gg \varphi$ and $\bigwedge_{c_\theta \in \mathsf{sub}(c_\varphi)} \theta = \varphi$, we have (25) iff

$$\exists v \leq w. \forall u \leq v : u \models c_\varphi : \varphi \ . \tag{26}$$

But this means $w \models B(c_\varphi : \varphi)$, which is equivalent by Lemma 4.8 to the statement that $w \models B(t:\varphi)$ for some term t.

For explicit belief, $w \models B^e\varphi$ means $w \models B(Ec_\varphi)$, which means

$$\exists v \leq w. \forall u \leq v : u \models Ec_\varphi \ . \tag{27}$$

Since M is a model and therefore satisfies Evidential Goodness, we have (27) iff

$$\exists v \leq w. \forall u \leq v : u \models Ec_\varphi \wedge \varphi \ . \tag{28}$$

Since $c_\varphi \gg \varphi$ and $\bigwedge_{c_\theta \in \mathsf{sub}(c_\varphi)} \theta = \varphi$, we have (28) iff

$$\exists v \leq w. \forall u \leq v : u \models Ec_\varphi \wedge c_\varphi : \varphi \ . \tag{29}$$

But (29) is what it means to have

$$\exists v \leq w. \forall u \leq v : u \models c_\varphi :^e \varphi \ . \tag{30}$$

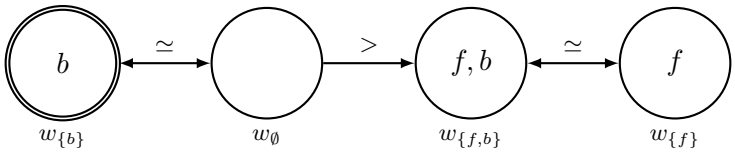
But this means $w \models B(c_\varphi :^e \varphi)$, which, by Lemma 4.8, is equivalent to the statement that $w \models B(t :^e \varphi)$ for some term t.

For $X = \Box$ and $X = \Box^e$, change the quantifiers "$\exists v \leq w. \forall u \leq v$" to "$\forall v \leq w$," change the B's to $\Box$'s, and change the B^e's to $\Box^e$'s. For $X = K$ and $X = K^e$, change the quantifiers to "$\forall v \sim w$," change the B's to K's, and change the B^e's to K^e's. □

5 Examples

5.1 The Gettier Example: Explicitly Justified True Belief Without Knowledge

This example is due to Edmund Gettier [32]. Our (anonymous) agent has a justified belief that Jones owns a Ford: indeed, she has seen Jones driving a Ford, Jones told her that it is his car and showed her the car ownership papers, etc. Let us denote by f that statement that "Jones owns a Ford" and by c_f the totality of the evidence in favor of f possessed by our agent: c_f looks pretty conclusive, and indeed if it were legitimate, then it would certainly support f. So our agent accepts this evidence (believes it to be good): $B(Ec_f)$ (i.e., $B^e f$). However, unknown to the agent, her



$$
\begin{aligned}
E(w_{\{b\}}) &:= \{c_\top, c_{f \Rightarrow b \vee f}\} \\
E(w_\emptyset) &:= E(w_{\{b\}}) \\
E(w_{\{f,b\}}) &:= \{c_\top, c_f, c_{f \Rightarrow b \vee f}, c_{f \Rightarrow b \vee f} \cdot c_f, c_{b \vee f}\} \\
E(w_{\{f\}}) &:= E(w_{\{f,b\}})
\end{aligned}
$$

Figure 1: Model for the Gettier Example (§5.1)

evidence c_f is *not* legitimate: the car papers are fake and in fact Jones does not own a Ford (but only borrowed one from a friend). So (in the actual world) the evidence c_f is *not* conclusive and f is in fact false.

Let b denote the sentence "Brown is in Barcelona." By conscious logical inference, our agent comes to the (explicit) justified belief that "Jones owns a Ford or Brown is in Barcelona": we have $B^e(b \vee f)$. But she has no evidence whatsoever about Brown: her justification for believing $b \vee f$ is only based on her belief about Jones.

In reality, unknown to the agent, Brown really *is* in Barcelona, so her belief in $b \vee f$ turns out to be true (i.e., correct). But is this justified true belief "knowledge"?

Formalization and Analysis of the Gettier Example

A model for this situation is pictured in Figure 1. Here we have four possible worlds, one world w_A for each $A \subseteq \{b, f\}$, with the obvious valuation for atomic sentences $p \in \{b, f\} =: \Phi$ (given by $w_A \models p$ iff $p \in A$). The actual world is $w_{\{b\}}$ (since Brown is in Barcelona and Jones does not own a Ford). Since the agent believes f, her most plausible worlds are $w_{\{f,b\}}$ and $w_{\{f\}}$, while the worlds $w_{\{b\}}$ and $w_\emptyset$ are strictly less plausible. Further, since she has no evidence for or against b, she considers the worlds $w_{\{f,b\}}$ and $w_{\{f\}}$ to be equally plausible, and the same goes for the worlds $w_{\{b\}}$ and $w_\emptyset$. To formalize the agent's explicit reasoning, we define E as in Figure 1 (where $c_\top$ is included in order to satisfy Trivial Evidence, Definition 3.1).

One can check that, in the actual world $w_{\{b\}}$, the following statements hold: $b \vee f$, $B^e f$, $B^e(b \vee f)$, $B((c_{f \Rightarrow b \vee f} \cdot c_f) :^e (b \vee f))$. So our agent has an explicitly justified true belief in $b \vee f$ based on the argument $c_{f \Rightarrow b \vee f} \cdot c_f$. However, it is easy to see that this is *not* "knowledge" according to our definition. Indeed, we have $\neg \Box^e (b \vee f)$, and moreover we have $\neg \Box (b \vee f)$: the agent's explicit belief in $b \vee f$ *is not explicit knowledge*, and in fact the agent *does not even have implicit knowledge* of $b \vee f$. This is confirmed by the defeasibility analysis: the actual world satisfies $\neg f \wedge \neg B^e(f \vee b | \neg f) \wedge \neg B(f \vee b | \neg f)$. In other words, if our agent would learn that Jones in fact does *not* own a Ford ($\neg f$), then she will lose (both her explicit and her implicit) belief in $f \vee b$.

5.2 The New Theorem Example: Explicitly Justified True Belief, and Implicit Knowledge, Without Explicit Knowledge

Our agent has a justified belief that a complicated JBG-formula θ consisting of more than 2^{25} symbols is true. Indeed, she has the testimony of Professor X: after showing her a multi-page printout of θ, the professor announces that θ is true.

Our agent's belief in θ is justified by her belief in the unfailing reliability of Professor X's logic-related pronouncements. We assume that the latter belief is supported by the testimony of several of our agent's logic-knowledgable friends, who have recommended the professor to our agent on this basis of their own belief in this unfailing reliability.

As it turns out, our agent's belief in θ is correct: θ is indeed true because it is a validity, derivable in JBG. However, unknown both to our agent and to her logic-knowledgeable friends, Professor X's logic-related pronouncements are not fully reliable: the professor often aims to test the students' knowledge and credulity by announcing the truth of complicated but randomly chosen formulas in the hope that the students will take this as a challenge to prove or disprove them. In particular, the professor's pronouncement with respect to θ is of this kind. Furthermore, he has not yet determined for himself whether θ is true, and has no idea whatsoever whether it is.

Our agent is a good student, and in fact she happens to know the axioms and rules of JBG needed to derive θ, but she has no other evidence pertaining to θ. In fact, she has never read or heard anything at all about this formula outside of her encounter with the professor, and she has not come across this formula in her own private thinking or study.

We let a denote the statement "Professor X announced θ to be true," and we let r denote the statement "Professor X's logic-related pronouncements are unfailingly reliable." The lattermost statement means that the professor only announces true statements about logic.

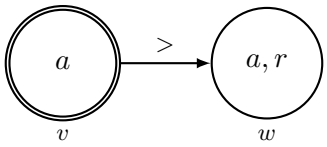
Our agent has direct perceptual evidence c_a in favor of a (since she heard the professor state that θ is true) and has some evidence c_r in favor of r (based on the testimony of her logic-knowledgable friends). She thus accepts the pieces of evidence c_a and c_r. Based on commonsense, she also knows that $a \Rightarrow (r \Rightarrow \theta)$; that is, if the professor announced that θ is true and his logic-related pronouncements are unfailingly reliable, then θ must be true. Let us denote by $c_{a\Rightarrow(r\Rightarrow\theta)}$ this commonsense justification. Using basic logic, our agent can construct a justification for her belief in θ in terms of her evidence c_a, c_r, and $c_{a\Rightarrow(r\Rightarrow\theta)}$. But is this explicitly justified true belief "knowledge"?

Formalization and Analysis of the New Theorem Example

We assume that our set Φ of propositional letters contains the symbols a and r in addition to the letters occurring in θ. As before, a says that Professor X announced θ is true, and r says that the professor's logic-related pronouncements are unfailingly reliable.

Since θ is a JBG-theorem, it follows by Internalization (Theorem 4.7) that there is a logical term $s \in \mathscr{T}$ such that $\mathsf{JBG} \vdash s : \theta$.

Logical terms are special because each is admissible for a unique JBG-theorem: a logical certificate c_φ is admissible only for the necessitated axiom φ (a JBG-theorem), and a logical term $t \cdot s$ is admissible only for a JBG-theorem ψ obtained by Modus Ponens (MP) from the unique JBG-theorem $\varphi \Rightarrow \psi$ for which t is admissible and the unique JBG-theorem φ for which s is admissible. In addition, the structure of a logical term t describes the structure of a specific JBG-derivation that begins with necessitated axioms and, using Modus Ponens as the only rule of inference, ends



$$\begin{aligned} E(v) &:= \mathsf{A} \cup \{c_\top, c_a, c_{a\Rightarrow(r\Rightarrow\theta)}, c_{a\Rightarrow(r\Rightarrow\theta)} \cdot c_a, c_{r\Rightarrow\theta}\} \\ E(w) &:= E(v) \cup \{c_r, c_{r\Rightarrow\theta} \cdot c_r, c_\theta\} \end{aligned}$$

(A is finite and $\mathsf{A}_s \subseteq \mathsf{A} \subseteq \mathsf{A}_{\mathsf{JBG}}$; see §5.2)

Figure 2: Model for the New Theorem Example (§5.2)

with the formula for which t is admissible. By Necessitation Elimination (Lemma 4.6), every JBG-theorem can be derived in this way. Accordingly, letting $\mathsf{A}_{\mathsf{JBG}}$ be the set of logical certificates, if we make explicitly available to the agent all members of the set

$$\mathsf{A}_s := \{c_\varphi \in \mathsf{A}_{\mathsf{JBG}} \mid c_\varphi \in \mathsf{sub}(s)\}$$

of logical certificates that make up s, then, since $s \gg \theta$, the agent has enough information to derive θ. Namely, she could obtain a derivation of θ, were she to turn her mind to this task, that starts from axioms whose certificates are available to her and proceeds via Modus Ponens as described by s. It is in this sense that we can realize our assumption that "the agent knows the axioms and rules of JBG needed to derive θ" by making the members of A_s explicitly available to the agent. However, since we assume that she knows nothing about θ other than what the professor told her—indicating that she has not turned her mind to finding a derivation of θ—we will assume that s is not explicitly available to her.

Our model for the present scenario, pictured in Figure 2, consists of two worlds v and w, with the valuation satisfying $[\![a]\!] = \{v, w\}$ and $[\![r]\!] = \{w\}$. Note that $[\![\theta]\!] = \{v, w\}$ because θ is valid. The actual world in Figure 2 is v (at which r is false, since Professor X's pronouncements are not unfailingly reliable), but our agent considers world w to be more plausible than v, so she mistakenly believes r. To represent our agent's reasoning, we define E as in Figure 2 (where $c_\top$ is included in order to satisfy Trivial Evidence, Definition 3.1). Since the agent at least knows all the axioms schemes and rules of JBG that are needed to derive θ, we assume that the contents of a finite set A satisfying $\mathsf{A}_s \subseteq \mathsf{A} \subseteq \mathsf{A}_{\mathsf{JBG}}$ are explicitly known to her.

One can check that at the actual world v the following statements hold: $B^e a$, $B^e r$, $B^e(r \Rightarrow \theta)$, $B((c_{a\Rightarrow(r\Rightarrow\theta)} \cdot c_a) :^e (r \Rightarrow \theta))$, $B((c_{r\Rightarrow\theta} \cdot c_r) :^e \theta)$, $B^e\theta$, and θ. In other words, the agent has constructed a chain of evidence that provides an explicit justification for her explicit belief in θ, and moreover this belief is correct.

Furthermore, unlike in the Gettier Example, the agent of the present example *has implicit defeasible knowledge of* θ. Indeed, $\Box\theta$ holds in the real world v because θ holds in *all* worlds: θ is after all a validity. In fact, since our agent knows the axiom schemes and rules of JBG needed to derive θ, the formula θ *really follows* from evidence that she actually possesses: the logical term s (describing a correct proof of θ) really does support θ, and s is constructible (in principle) using certificates $c_\varphi \in \mathsf{A}$ that are all in the agent's possession.

Nevertheless, the agent still *does not have explicit knowledge of* θ; that is, we have $v \not\models \Box^e\theta$.

This is because she does not actually possess the argument s; that is, $s \notin E(v) \cup E(w)$. The fact that her explicit justified belief in θ is not explicit knowledge is witnessed by the fact that we have $v \not\models B^e(\theta|\neg r)$. In words: if she were to learn that the professor's logic-related pronouncements are not unfailingly reliable, then she would lose her explicit belief that θ holds. (Though since she does have implicit knowledge that θ holds, she would nevertheless continue to have an implicit belief in it; that is, $v \models B(\theta|\neg r)$.)

6 Omniscience

Formal models of epistemic and doxastic logic often ascribe "too much" knowledge (or belief) to agents.[11] Whether an ascription of knowledge or belief is "too much" depends on one's philosophical views. To list just a few objections one might raise, let us associate with a pre-model (M, w) and a propositional attitude symbol $X \in \{B, B^e, \Box, \Box^e, K, K^e\}$ the set

$$X(M, w) := \{\varphi \in \mathscr{F} \mid w \models_M X\varphi\}$$

of formulas about which the agent has the attitude denoted by X. As before, B is implicit belief, B^e is explicit belief, $\Box$ is implicit (defeasible) knowledge, $\Box^e$ is explicit (defeasible) knowledge, K is implicit possession of information, and K^e is explicit possession of information. Here are just a few potentially worrying properties that the agent's knowledge or belief may satisfy.

- *(Sentential) Logical Omniscience* (see, e.g., [27, pp. 310–313]): the agent's knowledge or belief (set) is closed under a logical principle P. (Examples of P: "the agent knows all valid formulas," or "if the agent knows both φ and $\varphi \Rightarrow \psi$, then the agent also knows ψ.")

  This might be worrisome because "real agents" (e.g., humans, animals, and computers), as opposed to the "formal agents" we talk about for most of this paper, are generally logically non-omniscient: they (for a number of reasons) generally do not turn their minds to all possible logical consequences of their knowledge or beliefs and therefore cannot reasonably be construed to know or believe all such consequences.

- *(Sentential) Algorithmic Omniscience* (see, e.g., [35]): the agent's knowledge or belief (set) is not computable.[12]

  This might be worrisome for someone who holds the view that "real agents" cannot know something that cannot be computed by a Turing Machine.

- *(Sentential) Computational Omniscience* (see, e.g., [11, 12]): the agent's knowledge or belief (set) includes a validity φ whose shortest derivation in a given theory T exceeds some computational bound defined in terms of the size of φ (e.g., the shortest T-derivation of φ is not polynomial-time computable in the number of symbols that make up φ).

[11] We borrow this formulation from Melvin Fitting [28, 31].

[12] Example: taking the set $\Phi := \{p_k \mid k \in \mathbb{N}\}$ of propositional letters and fixing a bijection $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, define the standard model $M := (W, [\![\cdot]\!], \sim, \geq, E)$ by setting $W := \{w\}$, $w \in [\![p_k]\!]$ iff $f^{-1}(k) = (i, j)$ and the i-th Turing Machine halts on the j-th input (for some fixed language), $w \sim w$, $w \geq w$, and $E(w) := \{c_\top\} \cup \{c_{p_k} \mid w \in V(p_k)\}$. Then neither $K^e(M, w)$ nor $K(M, w)$ is computable: for each of these sets, were there a Turing Machine that halts on all inputs $\varphi \in \mathscr{F}$ and that accepts if and only if φ is in the set in question, then we would obtain a decision procedure for the Halting Problem.

This might be worrisome for someone who holds the view that "real agents" who reason in accordance with T (in some suitable sense) cannot perform computations that fall outside of a given class of "reasonably sized" computational problems.

- *(Sentential) Non-finitism* (see, e.g., [16, 26, 41, 51]): the agent's knowledge or belief (set) is infinite.

  This might be we worrisome for someone who holds the view that "real agents" (in light of their finite existence, finite memory, or finite computation or "thinking" speed, or for other reasons) cannot arrive in a state where they have gone to the trouble to come to know an infinite number of things.

- *(Sentential) Non-ultrafinitism* (see, e.g., [51, 55]): the agent's knowledge or belief (set) is not limited by a fixed finite bound N (e.g., the number of subatomic particles in the universe).

  This might be worrisome for someone who holds the view that "real agents" (in light of their finite resources or in light of fundamental limitations in the universe) have a fixed fundamental informational limitation.

In our framework, implicit attitudes of belief, knowledge, and information suffer from all of these properties. However, our *explicit* notions can be controlled (by careful definition of the good-evidence availability function E) to avoid each of these properties if it is so desired.[13] For example, one might reasonably argue that our model for the Gettier Example (Figure 1) avoids omniscience that is logical (at least for P's that imply all validities are known or believed), algorithmic, or computational in nature; further, this model certainly satisfies both finitism and ultrafinitism. Our model for the New Theorem Example (Figure 2), on the other hand, may be problematic when it comes to ultrafinitism; however, it might reasonably be argued that it avoids the other properties we have mentioned.

Our overall point is the following: our framework, similar to "awareness structures" in general [27, 36], is sufficiently flexible to address a wide range of philosophical views on omniscience. However, unlike simple awareness structures, our framework not only addresses omniscience but does so from the perspective of explicit evidence-based reasoning. Our work therefore contributes to the general study of logical omniscience in Justification Logic [11, 12, 28, 31, 42, 43, 52, 59, 60]. In particular, we suggest a "new" kind of omniscience:

- *Justification (or Proof) Omniscience*: the evidence (set $E(w)$) actually available to the agent (at a world w) satisfies a property P of evidence (sets) that is analogous to a property P' of knowledge or belief (sets) used in defining a particular notion of Sentential Omniscience (such as the notions listed above). Justification Omniscience therefore comes in various versions depending on the choice of property P. Examples: P is the property that "$E(w)$ is closed under any correct application of the Modus Ponens operation $s \cdot t$," P is the property that "$E(w)$ is not computable," P is the property that "$E(w)$ is not finite," or P is the property that "$E(w)$ is not ultrafinite."

[13] In essence, one may think of E as an "awareness function" in the sense of [27]: at world w, the agent is only "aware of" (and can only explicitly know or believe) the formulas $\varphi \in \mathscr{L}$ that are evidenced by some term $t \in E(w)$. However, unlike the "awareness sets" from [27], our "awareness sets" $E(w)$ also provide explicit justifications for those formulas the agent explicitly knows or believes at w, and these justifications have definite argumentative structure.

Justification Omniscience might be worrisome for someone who holds the view that the range of justifications that are actually available to "real agents" is limited for a reason corresponding to P (and is essentially the same reason for finding P'-Sentential Omniscience to be worrisome). For example, if P is the property that "$E(w)$ is not finite" (so that P' is the property that "a knowledge or belief set is not finite"), then the worry with P-Justification Omniscience, just as with P'-Sentential Omniscience (i.e., Non-finitism), is that "real agents" (in light of their finite existence, finite memory, or finite computation or "thinking" speed, or for other reasons) cannot arrive in a state where they have gone to the trouble to come to know an infinite number of things (where "things" means "justifications" in the case of P-Justification Omniscience and "sentences" in the case of P'-Sentential Omniscience).

Our models can be made (via suitable restrictions on the good-evidence availability map E) to realize (or violate) variants of justification non-omniscience.

We note that the underlying intuitive reason for the failure of a form of Sentential Omniscience is that real agents cannot do the work required to "derive all the justifications." This intuitive rationale for the failure of Sentential Omniscience forms the actual definition of Justification (or Proof) Omniscience. Therefore, in explaining the reason why a certain kind of Sentential Omniscience might be worrisome, it is natural to posit a specific kind of Justification (or Proof) Omniscience as the underlying rationale: the reason why it is worrisome to think that the agent can know or believe all the sentences satisfying a sentential property P' (i.e., that she is P'-Sententially Omniscient) is that it is worrisome to think that she can "form all the justifications" satisfying the corresponding evidential property P (i.e., that she is P-Justification Omniscient).

Finally, we note that while "awareness structures" [27, 36] can express all forms of Sentential Omniscience, they cannot express Justification Omniscience. For this, something like our good-evidence availability map E is needed.

7 Conclusion and Relationship with Previous Work

We have presented a complete, decidable logic of justifiable belief, defeasible knowledge, and explicit justification based on conclusive evidence. We have explained and demonstrated by way of example how this logic can be used to reason about Gettier-type situations [32], while addressing various kinds of omniscience.

This paper is closely connected with the authors' previous work [14], in which we studied similar evidence-based notions but in a framework that did not require the notion of conclusive evidence. In addition to imposing Evidential Goodness, we have also simplified matters here by leaving out the dynamics and focusing on the single-agent case. This allows us to highlight at the most basic level the way in which we adapt the evidential reasoning mechanisms of Justification Logic [10] to develop a particular theory of evidence-based (static) belief revision. (See [14] and the references therein for a discussion of the difference between static and dynamic notions of belief revision.)

A natural next direction for work is to study evidence-based informational events along the lines the authors studied previously [14] except with the condition of Evidential Goodness in place. Another possibility is to look to connections with recent work by van Benthem and Pacuit [58] studying a theory of "evidence management" that provides a different, semantic approach to evidence that nevertheless may have natural connections with our work here. Yet another possibility is to look to possible connections with the Justification Logic-based belief revision work of Kuznets

and Studer [46], which provides a new account of the connection between public announcements and evidence introduction.

Finally, while the notion of evidence considered here has a strong connection with proofs (as is the case for Justification Logics in general [3, 10]), there are other natural notions of evidence that one may consider, and these suggest other operations on evidence beyond the proof-related ones we considered here. In this sense, we are currently working on a number of promising ideas, which were all anticipated in the comments of one of our anonymous referees. The first is to expand our setting by connecting it to the philosophical literature that uses default logic and non-monotonic logics to deal with doxastic justification (see, e.g., [37, 38]). For instance, the certificates c_φ may be interpreted as defaults, supporting the conclusion φ only in "normal" circumstances. Alternatively, in addition to Artemov's operation $t \cdot s$ (which builds complex proofs by applying Modus Ponens with respect to classical implication), one may consider a new operation $t \otimes s$ that builds complex doxastic justifications by applying Modus Ponens with respect to the (non-monotonic) doxastic conditionals (as captured in our logic by conditional beliefs). Another idea is to connect to Bayesian epistemology [39] by providing a probabilistic semantics for justification. This suggests even more operations on evidence that correspond to more quantitative forms of evidence management. We think both these lines of research will be useful in analyzing the so-called "No False Lemma" Gettier counterexamples [40], which are currently outside of the scope of our setting.

Acknowledgements

We thank Professor Fitting for his editorial work, and we thank two anonymous referees for their valuable comments.

Dedication

This paper is dedicated to Professor Sergei Artemov on the occasion of his 60th birthday.

References

[1] E. Antonakos. Justified and common knowledge: Limited conservativity. In S. N. Artemov and A. Nerode, editors, *Logical Foundations of Computer Science, International Symposium, LFCS 2007, New York, NY, USA, June 4–7, 2007, Proceedings*, volume 4514 of *Lecture Notes in Computer Science*, pages 1–11. Springer, 2007.

[2] H. Arló-Costa and K. Kishida. Three proofs and the Knower in the Quantified Logic of Proofs. In *Online Proceedings of Sixth Annual Formal Epistemology Workshop (FEW 2009)*, Carnegie Mellon University, Pittsburg, PA, USA, June 18–21, 2009. Comments to [23].

[3] S. N. Artemov. Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, 7(1):1–36, Mar. 2001.

[4] S. N. Artemov. Justified common knowledge. *Theoretical Computer Science*, 357(1–3):4–22, July 2006.

[5] S. N. Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4):477–513, Dec. 2008.

[6] S. N. Artemov. Justification of knowledge: Philosophy and logic. In X. Arrazola and M. Ponte, editors, *LogKCA-10, Proceedings of the Second ILCLI International Workshop on Logic and Philosophy of Knowledge, Communication and Action*, pages 17–36. University of the Basque Country Press, 2010.

[7] S. N. Artemov. Tracking evidence. In A. Blass, N. Dershowitz, and W. Reisig, editors, *Fields of Logic and Computation, Essays Dedicated to Yuri Gurevich on the Occasion of His 70th Birthday*, volume 6300 of *Lecture Notes in Computer Science*, pages 61–74. Springer, 2010.

[8] S. N. Artemov. Why do we need Justification Logic? In J. van Benthem, A. Gupta, and E. Pacuit, editors, *Games, Norms and Reasons: Logic at the Crossroads*, volume 353 of *Synthese Library*, chapter 2, pages 23–38. Springer, 2011.

[9] S. N. Artemov. The ontology of justifications in the logical setting. *Studia Logica*, 100(1–2):17–30, Apr. 2012. Published online February 2012.

[10] S. N. Artemov and M. Fitting. Justification logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2012 edition, 2012.

[11] S. N. Artemov and R. Kuznets. Logical omniscience via proof complexity. In Z. Ésik, editor, *Computer Science Logic, 20th International Workshop, CSL 2006, 15th Annual Conference of the EACSL, Szeged, Hungary, September 25–29, 2006, Proceedings*, volume 4207 of *Lecture Notes in Computer Science*, pages 135–149. Springer, 2006.

[12] S. N. Artemov and R. Kuznets. Logical omniscience as a computational complexity problem. In A. Heifetz, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Twelfth Conference (TARK 2009)*, pages 14–23, Stanford University, California, July 6–8, 2009. ACM.

[13] S. N. Artemov and E. Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, Dec. 2005.

[14] A. Baltag, B. Renne, and S. Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In L. Ong and R. de Queiroz, editors, *Proceedings of the 19th Workshop on Logic, Language, Information, and Computation (WoLLIC 2012)*, volume 7456 of *Lecture Notes in Computer Science*, pages 168–190, Buenos Aires, Argentina, 2012. Springer-Verlag Berlin Heidelberg.

[15] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games, pages 9–58. Amsterdam University Press, 2008.

[16] P. Bernays. On Platonism in mathematics. In P. Benacerraf and H. Putnam, editors, *Philosophy of Mathematics: Selected Readings*, pages 258–271. Cambridge University Press, 2nd edition, 1983.

[17] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[18] S. Bucheli. *Justification Logics with Common Knowledge*. PhD thesis, Universität Bern, May 2012.

[19] S. Bucheli, R. Kuznets, B. Renne, J. Sack, and T. Studer. Justified belief change. In X. Arrazola and M. Ponte, editors, *LogKCA-10, Proceedings of the Second ILCLI International Workshop on Logic and Philosophy of Knowledge, Communication and Action*, pages 135–155. University of the Basque Country Press, 2010.

[20] S. Bucheli, R. Kuznets, and T. Studer. Two ways to common knowledge. In T. Bolander and T. Braüner, editors, *Proceedings of the 6th Workshop on Methods for Modalities (M4M–6 2009), Copenhagen, Denmark, 12–14 November 2009*, number 262 in Electronic Notes in Theoretical Computer Science, pages 83–98. Elsevier, May 2010.

[21] S. Bucheli, R. Kuznets, and T. Studer. Justifications for common knowledge. *Journal of Applied Non-Classical Logics*, 21(1):35–60, Jan.–Mar. 2011.

[22] B. F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

[23] W. Dean and H. Kurokawa. Knowledge, proof and the Knower. In *Online Proceedings of Sixth Annual Formal Epistemology Workshop (FEW 2009)*, Carnegie Mellon University, Pittsburg, PA, USA, June 18–21, 2009. Later version published as [24]; comments by H. Arló-Costa and K. Kishida available as [2].

[24] W. Dean and H. Kurokawa. Knowledge, proof and the Knower. In A. Heifetz, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Twelfth Conference (TARK 2009)*, pages 81–90, Stanford University, California, July 6–8, 2009. ACM.

[25] W. Dean and H. Kurokawa. From the Knowability Paradox to the existence of proofs. *Synthese*, 176(2):177–225, Sept. 2010. Published online May 2009.

[26] M. Dummett. Wang's paradox. *Synthese*, 30(3/4):301–324, 1975.

[27] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. The MIT Press, 1995.

[28] M. Fitting. A logic of explicit knowledge. In L. Běhounek and M. Bílková, editors, *Logica Yearbook 2004*, pages 11–22. Filosofia, Prague, 2005.

[29] M. Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005.

[30] M. Fitting. A quantified logic of evidence. *Annals of Pure and Applied Logic*, 152(1–3):67–83, Mar. 2008.

[31] M. Fitting. Reasoning with justifications. In D. Makinson, J. Malinowski, and H. Wansing, editors, *Towards Mathematical Philosophy, Papers from the Studia Logica conference Trends in Logic IV*, volume 28 of *Trends in Logic*, chapter 6, pages 107–123. Springer, 2009. Published online November 2008.

[32] E. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[33] M. Ghari. Distributed knowledge with justifications. In D. Lassiter and M. Slavkovik, editors, *New Directions in Logic, Language and Computation, ESSLLI 2010 and ESSLLI 2011 Student Sessions, Selected Papers*, volume 7415 of *Lecture Notes in Computer Science*, pages 91–108. Springer, 2012.

[34] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[35] J. Y. Halpern, Y. Moses, and M. Y. Vardi. Algorithmic knowledge. In *Proceedings of the 5th conference on Theoretical aspects of reasoning about knowledge*, pages 255–266. Morgan Kaufmann Publishers Inc., 1994.

[36] J. Y. Halpern and R. Pucella. Dealing with logical omniscience. In D. Samet, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the XIth conference (TARK 2007)*, pages 169–176, Brussels, Belgium, June 25–27, 2007. ACM.

[37] J. Horty. *Reasons as Defaults*. Oxford University Press, 2012.

[38] J. F. Horty. Reasons as defaults. *Philosophers' Imprint*, 7(3):1–28, 2007.

[39] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach.* Open Court Publishing Co, 1989.

[40] J. J. Ichikawa and M. Steup. The analysis of knowledge. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[41] D. Isles. What evidence is there that 2^65536 is a natural number? *Notre Dame Journal of Formal Logic*, 33(4), 1992.

[42] R. Kuznets. Complexity of evidence-based knowledge. In S. N. Artemov and R. Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge, 18th European Summer School in Logic, Language and Information (ESSLLI'06)*, pages 66–75, Málaga, Spain, Aug. 7–11, 2006. FoLLI.

[43] R. Kuznets. *Complexity Issues in Justification Logic*. PhD thesis, City University of New York, May 2008.

[44] R. Kuznets. Self-referentiality of justified knowledge. In E. A. Hirsch, A. A. Razborov, A. Semenov, and A. Slissenko, editors, *Computer Science — Theory and Applications, Third International Computer Science Symposium in Russia, CSR 2008, Moscow, Russia, June 7–12, 2008, Proceedings*, volume 5010 of *Lecture Notes in Computer Science*, pages 228–239. Springer, 2008.

[45] R. Kuznets. Self-referential justifications in epistemic logic. *Theory of Computing Systems*, 46(4):636–661, May 2010. Published online April 2009.

[46] R. Kuznets and T. Studer. Update as evidence: Belief expansion. In S. N. Artemov and A. Nerode, editors, *Proceedings of Logical Foundations of Computer Science LFCS'13*, volume 7734 of *Lecture Notes in Computer Science*, pages 266–279. Springer, 2013.

[47] K. Lehrer. *Theory of Knowledge*. Routledge, London, 1990.

[48] K. Lehrer. *Theory of Knowledge*. Westview Press, United States, 2000.

[49] K. Lehrer and T. J. Paxson. Knowledge: Undefeated justified true belief. *Journal of Philosophy*, 66:225–237, 1969.

[50] I. A. Letia and A. Groza. Arguing with justifications between collaborating agents. In P. McBurney, S. Parsons, and I. Rahwan, editors, *Argumentation in Multi-Agent Systems, 8th International Workshop, ArgMAS 2011, Taipei, Taiwan, May 3, 2011, Revised Selected Papers*, volume 7543 of *Lecture Notes in Artificial Intelligence*, pages 102–116. Springer, 2012.

[51] R. Parikh. Existence and feasibility in arithmetic. *The Journal of Symbolic Logic*, 36(3):494–508, 1971.

[52] R. Parikh. Logical omniscience and common knowledge; WHAT do we know and what do WE know? In R. van der Meyden, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Tenth Conference (TARK 2005)*, pages 62–77, Singapore, June 10–12, 2005. National University of Singapore.

[53] B. Renne. Public communication in justification logic. *Journal of Logic and Computation*, 21(6):1005–1034, Dec. 2011. Published online July 2010.

[54] B. Renne. Multi-agent justification logic: communication and evidence elimination. *Synthese*, 185(S1):43–82, Apr. 2012. Published online July 2011.

[55] V. Y. Sazonov. On feasible numbers. In D. Leivant, editor, *Logic and Computational Complexity*, volume 960 of *Lecture Notes in Computer Science*, pages 30–51. Springer Berlin Heidelberg, 1995.

[56] R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, number 2 in American Philosophical Quarterly Monograph Series, pages 98–112. Blackwell, 1968.

[57] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.

[58] J. van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99(1):61 – 92, 2011.

[59] R.-J. Wang. Knowledge, time, and logical omniscience. In H. Ono, M. Kanazawa, and R. de Queiroz, editors, *Logic, Language, Information and Computation, 16th International Workshop, WoLLIC 2009, Tokyo, Japan, June 21-24, 2009, Proceedings*, volume 5514 of *Lecture Notes in Artificial Intelligence*, pages 394–407. Springer, 2009. Later version published as [60].

[60] R.-J. Wang. Knowledge, time, and the problem of logical omniscience. *Fundamenta Informaticae*, 106(2–4):321–338, 2011.

[61] T. Yavorskaya (Sidon). Multi-agent explicit knowledge. In D. Grigoriev, J. Harrison, and E. A. Hirsch, editors, *Computer Science — Theory and Applications, First International Computer*

Science Symposium in Russia, CSR 2006, St. Petersburg, Russia, June 8–12, 2006, Proceedings, volume 3967 of *Lecture Notes in Computer Science*, pages 369–380. Springer, 2006. Journal version published as [62].

[62] T. Yavorskaya (Sidon). Interacting explicit evidence systems. *Theory of Computing Systems*, 43(2):272–293, Aug. 2008. Published online October 2007.