1

Why Justification Logic?

The formal details of justification logic will be presented starting with the next chapter, but first we give some background and motivation for why the subject was developed in the first place. We will see that it addresses, or at least partially addresses, many of the fundamental problems that have been found in epistemic logic over the years. We will also see in more detail how it relates to our understanding of intuitionistic logic. And finally, we will see how it can be used to mitigate some well-known issues that have arisen in philosophical investigations.

1.1 Epistemic Tradition

The properties of knowledge and belief have been a subject for formal logic at least since von Wright and Hintikka (Hintikka, 1962; von Wright, 1951). Knowledge and belief are both treated as modalities in a way that is now very familiar—*Epistemic Logic*. But of the celebrated three criteria for knowledge (usually attributed to Plato), *justified, true, belief*, Gettier (1963); Hendricks (2005), epistemic modal logic really works with only two of them. Possible worlds and indistinguishability model belief—one believes what is so under all circumstances thought possible. Factivity brings a trueness component into play—if something is not so in the actual world it cannot be known, only believed. But there is no representation for the justification condition. Nonetheless, the modal approach has been remarkably successful in permitting the development of rich mathematical theory and applications (Fagin et al., 1995; van Ditmarsch et al., 2007). Still, it is not the whole picture.

The modal approach to the logic of knowledge is, in a sense, built around the universal quantifier: X is known in a situation if X is true in *all* situations indistinguishable from that one. Justifications, on the other hand, bring an ex-

istential quantifier into the picture: X is known in a situation if *there exists* a justification for X in that situation. This universal/existential dichotomy is a familiar one to logicians—in formal logics there exists a proof for a formula X if and only if X is true in all models for the logic. One thinks of models as inherently nonconstructive, and proofs as constructive things. One will not go far wrong in thinking of justifications in general as much like mathematical proofs. Indeed, the first justification logic was explicitly designed to capture mathematical proofs in arithmetic, something that will be discussed later.

In justification logic, in addition to the category of formulas, there is a second category of *justifications*. Justifications are formal terms, built up from constants and variables using various operation symbols. Constants represent justifications for commonly accepted truths—axioms. Variables denote unspecified justifications. Different justification logics differ on which operations are allowed (and also in other ways too). If t is a justification term and X is a formula, $t{:}X$ is a formula, and is intended to be read

$$t \text{ is a justification for } X.$$

One operation, common to all justification logics, is *application*, written like multiplication. The idea is, if s is a justification for $A \to B$ and t is a justification for A, then $[s \cdot t]$ is a justification for B.[1] That is, the validity of the following is generally assumed

$$s{:}(A \to B) \to (t{:}A \to [s \cdot t]{:}B). \tag{1.1}$$

This is the explicit version of the usual distributivity of knowledge operators, and modal operators generally, across implication

$$\mathbf{K}(A \to B) \to (\mathbf{K}A \to \mathbf{K}B). \tag{1.2}$$

How adequately does the traditional modal form (1.2) embody epistemic closure? We argue that it does so poorly! In the classical logic context, (1.2) only claims that it is impossible to have both $\mathbf{K}(A \to B)$ and $\mathbf{K}A$ true, but $\mathbf{K}B$ false. However, because (1.2), unlike (1.1), does not specify dependencies between $\mathbf{K}(A \to B)$, $\mathbf{K}A$, and $\mathbf{K}B$, the purely modal formulation leaves room for a counterexample.

The distinction between (1.1) and (1.2) can be exploited in a discussion of the paradigmatic Red Barn Example of Goldman and Kripke; here is a simplified version of the story taken from Dretske (2005).

[1] For better readability brackets will be used in terms, "[,]", and parentheses in formulas, "(,)." Both will be avoided when it is safe.

> Suppose I am driving through a neighborhood in which, unbeknownst to me, papier-mâché barns are scattered, and I see that the object in front of me is a barn. Because I have barn-before-me percepts, I believe that the object in front of me is a barn. Our intuitions suggest that I fail to know barn. But now suppose that the neighborhood has no fake red barns, and I also notice that the object in front of me is red, so I know a red barn is there. This juxtaposition, being a red barn, which I know, entails there being a barn, which I do not, "is an embarrassment."

In the first formalization of the Red Barn Example, logical derivation will be performed in a basic modal logic in which $\Box$ is interpreted as the "belief" modality. Then some of the occurrences of $\Box$ will be externally interpreted as a knowledge modality $\mathbf{K}$ according to the problem's description. Let B be the sentence "the object in front of me is a barn," and let R be the sentence "the object in front of me is red."

(1) $\Box B$, "I believe that the object in front of me is a barn." At the metalevel, by the problem description, this is not knowledge, and we cannot claim $\mathbf{K}B$.
(2) $\Box(B \wedge R)$, "I believe that the object in front of me is a red barn." At the metalevel, this is actually knowledge, e.g., $\mathbf{K}(B \wedge R)$ holds.
(3) $\Box(B \wedge R \rightarrow B)$, a knowledge assertion of a logical axiom. This is obviously knowledge, i.e., $\mathbf{K}(B \wedge R \rightarrow B)$.

Within this formalization, it appears that epistemic closure in its modal form (1.2) is violated: $\mathbf{K}(B \wedge R)$, and $\mathbf{K}(B \wedge R \rightarrow B)$ hold, whereas, by (1), we cannot claim $\mathbf{K}B$. The modal language here does not seem to help resolving this issue.

Next consider the Red Barn Example in justification logic where $t{:}F$ is interpreted as "I *believe* F for reason t." Let u be a specific individual justification for belief that B, and v for belief that $B \wedge R$. In addition, let a be a justification for the logical truth $B \wedge R \rightarrow B$. Then the list of assumptions is

(i) $u{:}B$, "u is a reason to believe that the object in front of me is a barn";
(ii) $v{:}(B \wedge R)$, "v is a reason to believe that the object in front of me is a red barn";
(iii) $a{:}(B \wedge R \rightarrow B)$.

On the metalevel, the problem description states that (ii) and (iii) are cases of knowledge, and not merely belief, whereas (i) is belief, which is not knowledge. Here is how the formal reasoning goes:

(iv) $a{:}(B \wedge R \rightarrow B) \rightarrow (v{:}(B \wedge R) \rightarrow [a \cdot v]{:}B)$, by principle (1.1);
(v) $v{:}(B \wedge R) \rightarrow [a \cdot v]{:}B$, from 3 and 4, by propositional logic;
(vi) $[a \cdot v]{:}B$, from 2 and 5, by propositional logic.

Notice that conclusion (vi) is $[a \cdot v]{:}B$, and not $u{:}B$; epistemic closure holds. By reasoning in justification logic it was concluded that $[a \cdot v]{:}B$ is a case of knowledge, i.e., "I know B for reason $a \cdot v$." The fact that $u{:}B$ is not a case of knowledge does not spoil the closure principle because the latter claims knowledge specifically for $[a \cdot v]{:}B$. Hence after observing a red façade, I indeed know B, but this knowledge has nothing to do with (i), which remains a case of belief rather than of knowledge. The justification logic formalization represents the situation fairly.

Tracking justifications represents the structure of the Red Barn Example in a way that is not captured by traditional epistemic modal tools. The justification logic formalization models what seems to be happening in such a case; closure of knowledge under logical entailment is maintained even though "barn" is not perceptually known.

One could devise a formalization of the Red Barn Example in a bimodal language with distinct modalities for knowledge and belief. However, it seems that such a resolution must involve reproducing justification tracking arguments in a way that obscures, rather than reveals, the truth. Such a bimodal formalization would distinguish $u{:}B$ from $[a \cdot v]{:}B$ not because they have different reasons (which reflects the true epistemic structure of the problem), but rather because the former is labeled "belief" and the latter "knowledge." But what if one needs to keep track of a larger number of different unrelated reasons? By introducing a multiplicity of distinct modalities and then imposing various assumptions governing the interrelationships between these modalities, one would essentially end up with a reformulation of the language of justification logic itself (with distinct terms replaced by distinct modalities). This suggests that there may not be a satisfactory "halfway point" between a modal language and the language of justification logic, at least inasmuch as one tries to capture the essential structure of examples involving the deductive nature of knowledge.

1.2 Mathematical Logic Tradition

According to Brouwer, truth in constructive (intuitionistic) mathematics means the existence of a proof, cf. Troelstra and van Dalen (1988). In 1931–34, Heyting and Kolmogorov gave an informal description of the intended proof-based semantics for intuitionistic logic (Kolmogoroff, 1932; Heyting, 1934), which is now referred to as the *Brouwer–Heyting–Kolmogorov* (*BHK*) *semantics*. According to the *BHK* conditions, a formula is "true" if it hasa proof. Further-

more, a proof of a compound statement is connected to proofs of its components in the following way:

- a proof of $A \wedge B$ consists of a proof of proposition A and a proof of proposition B,
- a proof of $A \vee B$ is given by presenting either a proof of A or a proof of B,
- a proof of $A \rightarrow B$ is a construction transforming proofs of A into proofs of B,
- falsehood $\bot$ is a proposition, which has no proof; $\neg A$ is shorthand for $A \rightarrow \bot$.

This provides a remarkably useful informal way of understanding what is and what is not intuitionistically acceptable. For instance, consider the classical tautology $(P \vee Q) \leftrightarrow (P \vee (Q \wedge \neg P))$, where we understand $\leftrightarrow$ as mutual implication. And we understand $\neg P$ as $P \rightarrow \bot$, so that a proof of $\neg P$ would amount to a construction converting any proof of P into a proof of $\bot$. Because $\bot$ has no proof, this amounts to a proof that P has no proof—a refutation of P.

According to *BHK* semantics the implication from right to left in $(P \vee Q) \leftrightarrow (P \vee (Q \wedge \neg P))$ should be intuitionistically valid, by the following argument. Given a proof of $P \vee (Q \wedge \neg P)$ it must be that we are given a proof of one of the disjuncts. If it is P, we have a proof of one of $P \vee Q$. If it is $Q \wedge \neg P$, we have proofs of both conjuncts, hence a proof of Q, and hence again a proof of one of $P \vee Q$. Thus we may convert a proof of $P \vee (Q \wedge \neg P)$ into a proof of $P \vee Q$.

On the other hand, $(P \vee Q) \rightarrow (P \vee (Q \wedge \neg P))$ is not intuitionistically valid according to the *BHK* ideas. Suppose we are given a proof of $P \vee Q$. If we have a proof of the disjunct P, we have a proof of $P \vee Q$. But if we have a proof of Q, there is no reason to suppose we have a refutation of P, and so we cannot conclude we have a proof of $Q \wedge \neg P$, and things stop here.

Kolmogorov explicitly suggested that the proof-like objects in his interpretation ("problem solutions") came from classical mathematics (Kolmogoroff, 1932). Indeed, from a foundational point of view this reflects Kolmogorov's and Gödel's goal to define intuitionism within classical mathematics. From this standpoint, intuitionistic mathematics is not a substitute for classical mathematics, but helps to determine what is constructive in the latter.

The fundamental value of the *BHK* semantics for the justification logic project is that informally but unambiguously *BHK* suggests treating justifications, here mathematical proofs, as objects with operations.

In Gödel (1933), Gödel took the first step toward developing a rigorous proof-based semantics for intuitionism. Gödel considered the classical modal logic S4 to be a calculus describing properties of provability:

(1) *Axioms and rules of classical propositional logic*,
(2) $\Box(F \to G) \to (\Box F \to \Box G)$,
(3) $\Box F \to F$,
(4) $\Box F \to \Box\Box F$,
(5) *Rule of necessitation*: $\dfrac{\vdash F}{\vdash \Box F}$.

Based on Brouwer's understanding of logical truth as provability, Gödel defined a translation $tr(F)$ of the propositional formula F in the intuitionistic language into the language of classical modal logic: $tr(F)$ is obtained by prefixing every subformula of F with the provability modality $\Box$. Informally speaking, when the usual procedure of determining classical truth of a formula is applied to $tr(F)$, it will test the provability (not the truth) of each of F's subformulas, in agreement with Brouwer's ideas. From Gödel's results and the McKinsey-Tarski work on topological semantics for modal logic (McKinsey and Tarski, 1948), it follows that the translation $tr(F)$ provides a proper embedding of the Intuitionistic Propositional Calculus, IPC, into S4, i.e., an embedding of intuitionistic logic into classical logic extended by the provability operator.

$$\mathsf{IPC} \vdash F \quad \Leftrightarrow \quad \mathsf{S4} \vdash tr(F). \tag{1.3}$$

Conceptually, this defines IPC in S4.

Still, Gödel's original goal of defining intuitionistic logic in terms of classical provability was not reached because the connection of S4 to the usual mathematical notion of provability was not established. Moreover, Gödel noted that the straightforward idea of interpreting modality $\Box F$ as *F is provable in a given formal system $\mathcal{T}$* contradicted his second incompleteness theorem. Indeed, $\Box(\Box F \to F)$ can be derived in S4 by the rule of necessitation from the axiom $\Box F \to F$. On the other hand, interpreting modality $\Box$ as the predicate of formal provability in theory $\mathcal{T}$ and F as contradiction converts this formula into a false statement that the consistency of $\mathcal{T}$ is internally provable in $\mathcal{T}$.

The situation after Gödel (1933) can be described by the following figure where "$X \hookrightarrow Y$" should be read as "X is interpreted in Y":

$$\mathsf{IPC} \hookrightarrow \mathsf{S4} \hookrightarrow ? \hookrightarrow \textit{CLASSICAL PROOFS}.$$

In a public lecture in Vienna in 1938, Gödel observed that using the format of explicit proofs

$$t \textit{ is a proof of } F \tag{1.4}$$

can help in interpreting his provability calculus S4 (Gödel, 1938). Unfortunately, Gödel (1938) remained unpublished until 1995, by which time the

Gödelian logic of explicit proofs had already been rediscovered, axiomatized as the Logic of Proofs LP, and supplied with completeness theorems connecting it to both S4 and classical proofs (Artemov, 1995, 2001).

The Logic of Proofs LP became the first in the justification logic family. Proof terms in LP are nothing but *BHK* terms understood as classical proofs. With LP, propositional intuitionistic logic received the desired rigorous *BHK* semantics:

$$\mathsf{IPC} \hookrightarrow \mathsf{S4} \hookrightarrow \mathsf{LP} \hookrightarrow \textit{CLASSICAL PROOFS}\,.$$

Several well-known mathematical notions that appeared prior to justification logic have sometimes been perceived as related to the *BHK* idea: Kleene realizability (Troelstra, 1998), Curry–Howard isomorphism (Girard et al., 1989; Troelstra and Schwichtenberg, 1996), Kreisel–Goodman theory of constructions (Goodman, 1970; Kreisel, 1962, 1965), just to name a few. These interpretations have been very instrumental for understanding intuitionistic logic, though none of them qualifies as the *BHK* semantics.

Kleene realizability revealed a fundamental *computational content* of formal intuitionistic derivations; however it is still quite different from the intended *BHK* semantics. Kleene realizers are computational programs rather than proofs. The predicate "*r realizes F*" is not decidable, which leads to some serious deviations from intuitionistic logic. Kleene realizability is not adequate for the intuitionistic propositional calculus IPC. There are realizable propositional formulas not derivable in IPC (Rose, 1953).[2]

The Curry–Howard isomorphism transliterates natural derivations in IPC to typed λ-terms, thus providing a generic functional reading for logical derivations. However, the foundational value of this interpretation is limited because, as proof objects, Curry–Howard λ-terms denote nothing but derivations in IPC itself and thus yield a circular provability semantics for the latter.

An attempt to formalize the *BHK* semantics directly was made by Kreisel in his theory of constructions (Kreisel, 1962, 1965). The original variant of the theory was inconsistent; difficulties already occurred at the propositional level. In Goodman (1970) this was fixed by introducing a stratification of constructions into levels, which ruined the *BHK* character of this semantics. In particular, a proof of $A \to B$ was no longer a construction that could be applied to *any* proof of A.

[2] Kleene himself denied any connection of his realizability with the *BHK* interpretation.

1.3 Hyperintensionality

Justification logic offers a formal framework for hyperintensionality. The *hyperintensional paradox* was formulated in Cresswell (1975).

> It is well known that it seems possible to have a situation in which there are two propositions p and q which are logically equivalent and yet are such that a person may believe the one but not the other. If we regard a proposition as a set of possible worlds then two logically equivalent propositions will be identical, and so if "x believes that" is a genuine sentential functor, the situation described in the opening sentence could not arise. I call this the paradox of hyperintensional contexts. Hyperintensional contexts are simply contexts which do not respect logical equivalence.

Starting with Cresswell himself, several ways of dealing with this have been proposed. Generally, these involve adding more layers to familiar possible world approaches so that some way of distinguishing between logically equivalent sentences is available. Cresswell suggested that the syntactic form of sentences be taken into account. Justification logic, in effect, does this through its mechanism for handling justifications for sentences. Thus justification logic addresses some of the central issues of hyperintensionality but, as a bonus, we automatically have an appropriate proof theory, model theory, complexity estimates, and a broad variety of applications.

A good example of a hyperintensional context is the informal language used by mathematicians conversing with each other. Typically when a mathematician says he or she knows something, the understanding is that a proof is at hand, but this kind of knowledge is essentially hyperintensional. For instance Fermat's Last Theorem, FLT, is logically equivalent to $0 = 0$ because both are provable and hence denote the same proposition, as this is understood in modal logic. However, the context of proofs distinguishes them immediately because a proof of $0 = 0$ is not necessarily a proof of FLT, and vice versa. To formalize mathematical speech, the justification logic LP is a natural choice because $t{:}X$ was designed to have characteristics of *"t is a proof of X."*

The fact that propositions X and Y are equivalent in LP, that $\mathsf{LP} \vdash X \leftrightarrow Y$, does not warrant the equivalence of the corresponding justification assertions, and typically $t{:}X$ and $t{:}Y$ are not equivalent, $t{:}X \not\leftrightarrow t{:}Y$. Indeed, as we will see, this is the case for every justification logic.

Going further LP, and justification logic in general, is not only sufficiently refined to distinguish justification assertions for logically equivalent sentences, but it also provides flexible machinery to connect justifications of equivalent sentences and hence to maintain constructive closure properties desirable for a logic system. For example, let X and Y be provably equivalent, i.e., there is a proof u of $X \leftrightarrow Y$, and so $u{:}(X \leftrightarrow Y)$ is provable in LP. Suppose also

that v is a proof of X, and so $v{:}X$. It has already been mentioned that this does not mean v is a proof of Y—this is a hyperintensional context. However within the framework of justification logic, building on the proofs of X and of $X \leftrightarrow Y$, we can *construct* a proof term $f(u, v)$, which represents the proof of Y and so $f(u, v){:}Y$ is provable. In this respect, justification logic goes beyond Cresswell's expectations: Logically equivalent sentences display different but constructively controlled epistemic behavior.

1.4 Awareness

The logical omniscience problem is that in epistemic logics all tautologies are known and knowledge is closed under consequence, both of which are unreasonable. In Fagin and Halpern (1988) a simple mechanism for avoiding the problems was introduced. One adds to the usual Kripke model structure an *awareness* function $\mathcal{A}$ indicating for each world which formulas the agent is aware of at this world. Then a formula is taken to be known at a possible world u if (1) the formula is true at all worlds accessible from u (the Kripkean condition for knowledge) and (2) the agent is aware of the formula at u. The awareness function $\mathcal{A}$ can serve as a practical tool for blocking knowledge of an arbitrary set of formulas. However, as logical structures, awareness models exhibit abnormal behavior due to the lack of natural closure properties. For example, the agent can know $A \wedge A$ but be unaware of A and hence not know it.

Fitting models for justification logic, presented in Chapter 4, use a forcing definition reminiscent of the one from awareness models: For any given justification t, the justification assertion $t{:}F$ holds at world u iff (1) F holds at all worlds v accessible from u and (2) t is an admissible evidence for F at u, $u \in \mathcal{E}(s, F)$, read as "u is a possible world at which s is relevant evidence for F." The principal difference is that postulated operations on justifications relate to natural closure conditions on admissible evidence functions $\mathcal{E}$ in justification logic models. Indeed, this idea has been explored in Sedlár (2013), which works with the language of LP and thinks of it as a multiagent modal logic, and taking justification terms as agents (more properly, actions of agents). This shows that justification logic models absorb the usual epistemic themes of awareness, group agency, and dynamics in a natural way.

1.5 Paraconsistency

Justification logic offers a well-principled approach to paraconsistency, which looks for noncollapsing logical ways of dealing with contradictory sets of assumptions, e.g.,

$$\{A, \neg A\}.$$

The following obvious observation shows how to convert any set of assumptions

$$\Gamma = \{A_1, A_2, A_3, \ldots\}$$

into a logically consistent set of sentences while maintaining all the intrinsic structure of Γ. Informally, instead of (perhaps inconsistently) assuming that Γ holds, we assume only that each sentence A from Γ has a justification, i.e.,

$$\vec{x}{:}\,\Gamma = \{x_1{:}A_1, x_2{:}A_2, x_3{:}A_3, \ldots\}.$$

It is easy to see that for each Γ, the set $\vec{x}{:}\Gamma$ is consistent in what will be our basic justification logic J.

For example, for $\Gamma = \{A, \neg A\}$,

$$\vec{x}{:}\,\Gamma = \{x_1{:}A, x_2{:}\neg A\},$$

states that x_1 is a justification for A and x_2 is a justification for $\neg A$. Within justification logic J in which no factivity (or even consistency) of justifications is assumed, the set of assumptions $\{x_1{:}A, x_2{:}\neg A\}$ is consistent, unlike the original set of assumptions $\{A, \neg A\}$.

There is nothing paraconsistent, magical, or artificial in reasoning from $\vec{x}{:}\Gamma$ in justification logic J. In practical terms, this means we gain the ability to effectively reason about inconsistent data sets, keeping track of justifications and their dependencies, with the natural possibility to draw meaningful conclusions even when some assumed justifications from $\vec{x}{:}\Gamma$ become compromised and should be discharged.