Methods Models for Belief and Knowledge*

1st September 2010

FEW 2010 (Konstanz) draft
in progress, comments very welcome

julien.dutant@unige.ch

Abstract

We introduce a formal representation of belief and knowledge based on the idea that knowledge is a matter of forming a belief through a sufficiently error-free method. We first model methods and their infallibility, then define belief and knowledge in terms of them. The resulting models are a significant extension of so-called "neighbourhood models". We argue that epistemological notions and problems like Gettier, inductive knowledge, fallible justification, epistemic contextualism, or failure of logical omniscience are represented in a more satisfactory ways in these models than in standard epistemic logic. In general, our models only validate the claim that knowledge is true belief; but we show that a full **S5** system can be derived from a set of natural idealisations. The derivation provides some explanation of why and when the **S5** axioms should hold, and a vindication of their use.

*Thanks to Paul Egré, Elia Zardini, Jonathan Shaheen, Davide Fassio, Fabrice Correia, Olivier Roy, Andreas Witzel, Vincent Hendricks, Johan van Benthem, John Hawthorne, Jonas de Vuyst, Timothy Williamson for much-needed help and encouragement, and to audiences in Geneva (Palmyr workshop), Lausanne (DLG09), Nancy (EpiConfFor workshop) and two anonymous referees for useful comments.

Contents

1 Introduction

My starting point is the following puzzle. (1) Whether one knows centrally depends on what basis one's belief has. (2) Standard epistemic logic cannot represent bases of belief. (3) Standard epistemic logic adequately models knowledge in a number of applications. I introduce formal models of knowledge directly stemming from the idea that knowledge is a matter of bases of belief, or as I will call them, *methods*. The models are an extension of Scott's (1970) and Montague's (1968; 1970) neighbourhood models, and they differ from other recent formal systems aimed at dealing with similar issues.[1] They solve the puzzle by providing an insight into why and when axioms of standard epistemic logic hold, including the most controversial ones. But most importantly, they provide a philosophically satisfying representation of knowledge. We illustrate the point by using them to formalise several classical epistemological problems and views: the Gettier problem, inductive knowledge, deductive closure without logical omniscience, Frege cases, epistemic contextualism, and failure of knowledge to iterate.

Let me first illustrate the puzzle. Consider three cases:

1. *Tea leaves.* As things are, reading tea leaves is not a reliable way to find the truth. My uncle believes on the sole basis of tea leaves readings that I will get a pay raise soon. Whether or not I will, he does not know that I will.

2. *Watson.* Holmes and Watson know the same facts about a case. Reasoning carefully, Holmes deduces that the father is the culprit. Watson is also convinced of this, but on the sole basis of the father's shady looks. Watson does not know that the father did it.

3. *Induction.* Seeing that the light is on at the neighbour's, my mother infers that the neighbours are home. In suitably normal circumstances, she comes to know that the neighbours are home.

Here are a few *prima facie* intuitive things to say about the cases. In (1)-(2), my uncle and Watson fail to know because the bases of their beliefs are not adequate for knowledge. In Watson's case, that is so even though his belief is true and he knows facts which together entail it. In case (3), my mother comes

[1] Notably Kelly's (1996) Learning Theory, Artemov's Logic of Proofs (1994; 2005), Fagin and Halpern's (1988) Awareness Models. We will not provide a detailed comparison of the present models with these alternatives here.

to know something on a basis which fails to entail it. That is so because her basis is adequate or sufficient given the circumstances she is in. All three cases show that consideration of the basis of one's belief and its adequateness to the circumstances are central to whether one knows.

Given the apparently central role of bases of belief in epistemology, it is striking how hard it is to accommodate the notion in standard models for knowledge introduced by Hintikka (1962). Such models essentially characterise knowledge and belief in terms of *elimination of possibilities*:

> What the concept of knowledge involves in a purely logical perspective is thus a dichotomy of the space of all possible scenarios into those that are compatible with what I know and those that are incompatible with my knowledge. Hintikka (2007, 15)

As Hintikka makes clear, the account is not reductive, since the "elimination" of possibilities is itself defined in terms of knowledge.[2] In fact, standard epistemic logic is best construed as a representation of the *content* of one's knowledge rather than as a representation of the state of knowing.[3] That does not mean that it does not say anything about knowledge. The problem is rather that what it says seems false, namely that one knows p iff one knows something incompatible with $\neg p$. In our *Watson* case, Watson fails to know that the father is the culprit even though he knows things that are incompatible with it being false. In our *Induction* case, my mother appears to learn that the neighbours are home on the basis of facts compatible with them not being here. (One may of course reply that before she came to learn it, the facts known to her where indeed compatible with them not being there, but then she did not know they were there; while since she has drawn the inference, she knows some fact incompatible with them not being there: simply, that they *are* there. Be that as it may, we still lack an account of how an inductive inference can achieve

[2] Lewis (1996) attempts to turn the idea into a reductive account by construing elimination as metaphysical incompatibility with one's being in the total experiential state on is in. The move has unpalatable consequences (see e.g. Hawthorne, 2004, 60n).

[3] See Hintikka (2007, 16): "Epistemic logic presupposes essentially only the dichotomy between epistemically possible and epistemically excluded scenarios. How this dichotomy is drawn is a question pertaining to the definition of knowledge. However, we do not need to know this definition in doing epistemic logic. Thus the logic and the semantics of knowledge can be understood independently of any explicit definition of knowledge. Hence it should not be surprising to see that a similar semantics and a similar logic can be developed for other epistemic notions—for instance, belief, information, memory, and even perception. This is an instance of a general law holding for propositional attitudes. This law says that the content of a propositional attitude can be specified independently of differences between different attitudes."

such a result.) As we pointed out, a natural first step in thinking about these cases is to formulate them in terms of bases of belief or methods. But it is unclear how to introduce such a notion in standard models. In my view, that goes a long way towards explaining the widely noted gap between epistemic logic and epistemology (Hendricks, 2006; van Benthem, 2006): epistemologists mainly think of knowledge as adequately based belief, while epistemic logic represents it as the elimination of possibilities, and it is unclear how to fit the two pictures together.

I am not claiming that epistemic logic cannot be amended and fruitfully extended to deal with some methods-related aspects of knowledge. Much has already be done in that respect.[4] But I take a different approach here. Instead of adapting standard models to specific philosophical purposes, I start from a philosophical characterisation of knowledge and build a model suited to it.

Call *methods* whatever bases of beliefs or ways of forming or sustaining beliefs are relevant to knowledge attribution.[5] Believing that an object is an apple on the basis of sight at a short distance involves a different method than believing it on the basis of sight at a great distance; believing that a market is going to collapse on the basis of hearsay involves a different method than believing it on the basis of an expert report; believing that a man was present at a certain dinner on the basis of plausible inference involves a different method than believing it on the basis of a clear memory. Methods need not be conscious procedures one follows methodically. They may involve unconscious computational processes. They may be individuated "externally" or "broadly", that is, they may involve relations between an agent and her environment and aspects of the latter. We need not settle these important issues here. As will be become clear, a lot can be said about methods while maintaining a fairly abstract perspective on them, and it is sufficient for our purposes that our models are compatible with various substantive conceptions of what methods are.[6]

[4] See e.g. Fagin et al. (1995, chap. 9,10).

[5] That use of the term has some currency in epistemology since Nozick (1981, chap.3).

[6] I doubt that we can characterise methods in terms wholly independent of the notion of knowledge, such as psycho-physical descriptions. Rather, we have to use our judgements about knowlege as a guide to what methods are. For instance, we cannot count all beliefs based on sight at a short distance are based on a same method. For one may know that an object is a banana while mistaking an apple next to it for a peach, even though one sees both at a short distance. The two beliefs would thus have to involve different methods. Ultimately, the aim is to pick up methods through their relations to belief and knowledge and their structural properties identified by the models. But these points are beyond the scope of the present paper. (See Williamson (2000, 100) for analogous considerations with respect to the notion of safety used in an account of knowledge.)

(We make one important presupposition about methods, though. We assume that whether one knows in a given case depends on an equivalence class of beliefs with the *same* basis, and not on a class of beliefs with *similar enough* bases, where the similarity relation may be non-transitive. [7] We leave for further work the exploration of how our models may be recast in the less demanding setting of a similarity relation between bases.)

Our guiding idea is that knowledge is belief based on an infallible method:

Methods infallibilism An agent knows that p iff she believes that p on the basis of a method that could only yield true beliefs.

We say that a method "yields" a belief whenever it forms it or sustains it. A method that produces truth-valueless beliefs (if there are such) counts as fallible by the definition. The "could" is meant to be understood as a (possibly context-sensitive) alethic modal.

There are several reasons to adopt the method-infallibilist account, which we can only mention here. First, it automatically secures factivity and the idea that deduction preserves knowledge – which is to be distinguished from logical omniscience, as we will see. Second, analyses that allow adequate bases for belief to be compatible with error at the circumstances at hand seem to systematically face Gettier-style counterexamples; some infallibility condition appears required to avoid them (Sturgeon, 1993). Third, it provides a simple diagnosis of why no matter how high the odds, we fail to know in advance that a ticket in a fair lottery is a looser (Hawthorne, 2004). I also hope that the intuitive results we get from the formal implementation of the account will further support it. Be that as it may, the models we introduce can alternatively be extended to represent fallibilist notions of knowledge.

A crucial aspect of the methods approach is that in order to evaluate whether a particular belief is knowledge, we have to consider *other* beliefs one has or could have had on the same basis. Consider in particular:

1. *Fake oranges.*[8] Looking at a particular orange in a fruit bowl, Oscar believes that *that orange* is a fruit. Unbeknownst to him, the other "fruits" in the bowl are perfect wax replicas of oranges.

2. *Prime numbers.* Primo has a mistaken way of evaluating whether a number higher than 20 is prime: he adds its digits, and if the sum is prime he

[7] Thanks to John Hawthorne here.
[8] A variant of Ginet-Goldman barn facades case (Goldman, 1976, 772–3).

judges that the original number was too. For instance, since $4+7=11$ is prime, he (rightly) believes that 47 is prime.

Assuming essentialism, there is no possible world where *that* orange, the one Oscar is looking at, is not a fruit. And there is no possible worlds where 47 is not prime or where Primo's method would lead him to wrongly believe that it is not. Yet both fail to know, because they could have believed false propositions on the basis of the same methods: Oscar would falsely believe on the same basis that an "orange" next to the one he is looking at was a fruit too, and Primo would falsely believe on the same basis that 49 is prime. The fact that, so to speak, we evaluate a belief by looking at whether it is in "good" or "bad" company is the basic idea of our formalisation of methods.

Section 2 gives the most general characterisation of methods and infallibility, without assuming a specific notion of proposition. We define operations on methods and give an algebra for them. Section 3 applies the method-infallibilist approach to formalise Gettier cases, fallible justification and inductive knowledge. Section 4 introduces methods models properly. For concreteness and simplicity we take propositions to be sets of possible worlds. That creates trouble with so-called Frege cases, though not as straightforwardly as expected. The resulting models are an extension of neighbourhood models, but more explanatory than the latter. (A detailed comparison is made in appendix B, sec 8). We introduce a language for methods-based belief and knowledge. Section 5 details the main consequences of the models. In general, the models only validate the uncontroversial claim that knowledge is true belief; however, we derive a full **S5** system for a series of natural idealisations of the agent's methods. The derivation provides a illuminating perspective on why and when the axioms of standard epistemic logic hold.

2 Methods

2.1 The space of possible methods

A method is something that yields beliefs. But it does not need to yield the same beliefs wherever it exists; in fact, it typically yield beliefs as a function of other factors. First, a method may yield beliefs as a function of the state of the world. Facing a table, Alice opens her eyes. She immediately forms beliefs: say, that there is an apple, or that there is an apple and that there is a pear,

depending on what there is on the table. That is the idea of a *purely non-inferential* method, and it is naturally modelled as a function from worlds to sets of propositions. Second, a method may yield beliefs as a function of other beliefs. For instance, *modus ponens* would lead you to believe that q from the premises that p and that $p \rightarrow q$. That is the idea of a *purely inferential* method, and it is naturally modelled as a function from sets of propositions (premises) to sets of propositions (conclusions).

Generalising both ideas, a purely non-inferential method is a function from worlds and sets of premises that is constant over sets of premises. It yields a set of "conclusions" depending on the state of the world, for any premises whatsoever, including no premise at all. A purely inferential method is a function from worlds and sets of premises that is constant over worlds. It yields the same conclusions for a given set of premises, whatever world one is in. Mixed methods (if there are any) are variable functions from worlds and sets of premises to sets of conclusions.[9]

We are now in position to define the space of all possible methods as the space of all functions from worlds to functions from sets of premises to sets of conclusions. Let W be a set of worlds and P a set of propositions:

Definition 1. $\mathbf{M} = W \times (\mathcal{P}(P) \times \mathcal{P}(P))$ is the *space of possible methods*.

Terminology. For any method $m \in M$, world $w \in W$, and sets of propositions $\pi, \pi' \subseteq P$,
when $\pi' = m(w)(\pi)$, we say that π' is the *set of conclusions* reached by method m at world w from the set of premises π,
when $p \in m(w)(\pi)$, we say that p is a *conclusion* reached by m at w from π,
when $p \in m(w)(\emptyset)$, we say that p is an *unconditional output* of m at w, that is, a conclusion reached by m at t without premise.

Notation. We abbreviate:
$m(w, \pi) := m(w)(\pi)$,
$m(w) := m(w)(\emptyset)$.

Remark 1. By convention, worlds are noted w, w', ..., propositions p, p', ..., q, q', ..., sets of premises π, π', ..., and methods m, m', ..., n, n', ... We typically omit domains when they are clearly indicated by this convention. Thus we write: "for all w", "$\forall w(\ldots)$" and "$\{w : \ldots\}$" instead of "for all w in

[9] Here are two candidates cases of mixed methods: (1) perceptual processes that are sensitive to one's background beliefs; (2) trust processes (roughly, inferences from *S said p* to *p*) that are sensitive to subtle visual clues in a quasi-perceptual way. But whether or not one wants to contend such methods does not matter for the purposes of this paper.

W", "$\forall w \in W$" and "$\{w \in W : \ldots\}$", and similarly for any $p \in P$, $\pi \subseteq P$ and $m \in \mathbf{M}$.

Now the idea that methods are functions from *worlds* is a gross simplification. Suppose two agents are in distinct rooms, each facing a table: Alice sees an apple, Bob a pear. Intuitively, we would like to say that the very same method can produce a belief that there is a apple in Alice's mind, and no such belief in Bob's mind. Also, it is natural to take the outputs of a method to depend on the time and not just the world. To model these phenomena, methods should rather be functions from *centred worlds* $< c, w >$ where w is a world and c a *perspective* on that world. A perspective is a point from which methods can be applied; a time and a place, at least. Perspectives need not only be where some agent is: a visual method can be fallible in virtue of producing false beliefs as used from the top of the mountain, even though nobody has been, is or will be at the top of the mountain. The *life* of an agent is then the series of perspectives the agent occupies. The agent's belief and knowledge are then derived from her life and the methods she has.

These refinements for perspectives and lives are relatively straightforward to introduce, but they needlessly complicate the models for our present purposes. We thus stick to the characterisation of methods in terms of functions from worlds.

2.2 Operations on methods

For any methods m, n, we define:

Definition 2. The *union of* m *and* n is the method $(m + n)$ given by $(m + n)(w, \pi) := m(w, \pi) \cup n(w, \pi)$ for any w, π.

The *composition of* m *and* n is the method $(m \circ n)$ given by $(m \circ n)(w, \pi) := m(w, n(w, \pi))$ for any w, π.

Method union is the idea that an agent is able to pool together the outputs of different methods. If, given premises π, m outputs $\{p\}$ and n outputs $\{q\}$, the method $(m + n)$ outputs $\{p, q\}$. Putting limitations on union is thus a way to model the *modularity* of an agent. For instance, a limited number of unions corresponds to an agent with limited working memory, who is unable to put to use all her beliefs at once. And a systematic bar on uniting certain methods corresponds to an agent whose bodies of beliefs are partly isolated from each other.

Method composition is the idea that an agent is able to apply one method to the output of another: if n outputs $\{p\}$ from no premises, and m outputs $\{q\}$ from $\{p\}$, then $m \circ n$ outputs $\{q\}$ from no premises. Limits on an agent's ability to compose methods is thus a way to represent *computationally bounded* agents, such as agents who are only able to go through proofs with a limited number of steps.[10]

Given a set of methods M, we can define its union and composition closure as the smallest set $M^{\circ +}$ such that $M \subseteq M^{\circ +}$ and for any $m, n \in M^{\circ +}$, $(m + n), (m \circ n) \in M^{\circ +}$. $M^{\circ +}$ is the set of methods available to an agent that has the M methods and is neither modular nor computationally bounded.

Method union and composition are interpreted in a synchronic way here. If a (non-modular) agent has m and n at a given time, then $(m + n)$ is available to her at the very same time. However, union and composition could easily be construed as dynamic processes. For instance, given a certain method set M_0, one could consider the series $M_1, M_2, \ldots$ where each M_k corresponds to applying one step of union and composition to M_{k-1}: $M_k = M_{k-1} \cup \{(m+n) : m, n \in M_{k-1}\} \cup \{(m \circ n) : m, n \in M_{k-1}\}$ for $k \geq 1$. We will not get into such models here.

$\langle \mathbf{M}, +, \circ \rangle$ is an algebraic structure over the set of methods. We detail its main properties in Appendix A (section 7). The most remarkable are:

1. Method union is associative, commutative and idempotent:
   $(m + n) + r = m + (n + r)$.
   $m + n = n + m$.
   $m + m = m$.

2. Method composition is associative, but not commutative nor idempotent.
   $m \circ (n \circ r) = (m \circ n) \circ r$.

3. Composition distributes right-to-left over union, but not left-to-right:
   $(m + n) \circ r = (m \circ r) + (n \circ r)$,
   but $m \circ (n + r) = (m \circ n) + (m \circ r)$ may fail.

The fact that composition does not distribute left to right over union is a reflection of the fact that method composition keeps track of information processing or, more accurately, of information dependencies. $m \circ (n + r)$ corresponds to uniting n and r and *then* applying m, which is not the same as applying m

[10] I am grateful to Jonathan Shaheen and Andreas Witzel for helping me sorting out initial issues with modelling method composition with function composition.

to n and to r separately. To take a simple example: suppose m outputs the conjunction of any premises and suppose n outputs only p and r outputs only q. The union of n and r outputs both p and q, so $m \circ (n + r)$ will output $p \wedge q$. But applying m to the premise p and separately to the premise q will not yield the conjunction of the two, so $(m \circ n) + (m \circ r)$ will not output $p \wedge q$.

2.3 Infallibility

The infallibility of a method has two components. For its non-inferential part, it is infallible if it could only yield true beliefs. For its inferential part, it is infallible if it could only reach true conclusions from true premises. The fact that *modus ponens* reaches false conclusions from false premises does not make it infallible. In a nutshell: a method is infallible iff it could not reach false conclusions from any set of true premises, including the empty one.[11]

The relevant modality is alethic, not epistemic or doxastic. For a method to be infallible, it is not required that one knows or believes that it is. It is sufficient that error is in fact impossible. Within the domain of alethic modalities, many options are open. For instance, one could require that error be physically impossible given the makeup of the agent and the situation she is in (Armstrong, 1973, 168), or that error be impossible in sufficiently similar cases (Williamson, 2000, 100), or that error be impossible in a contextually determined set of relevant possibilities (Lewis, 1996, 553–4). We need not decide between them here. Three points should me mentioned, though.

Infallibility requires at least no error at the actual world: a method that actually yields a false belief is not such that it could not yield one. That is a consequence of the fact that the relevant modality is alethic.

Infallibility need not require impossibility of error across all worlds. In some worlds, pigs can fly, in others, they cannot. Similarly, some methods are fallible at some worlds and yet infallible at others.[12]

An important feature of the notion of possibility is whether what is possibly possible is possible (axiom **4** of modal logic). Physical possibility and related

[11] On a weaker notion of infallibility only the non-inferential component of methods (*i.e.* beliefs) would be taken into account. Thanks to Timothy Williamson for suggesting the stronger notion used here.

[12] In many applications of epistemic logic, one need only consider a restricted set of possible worlds — for instance, we simply ignore possible worlds in which our two prisoners communicate. In such set-ups, infallibility across "all" worlds is in effect a restricted form of infallibility. In other applications, the fact that a method can be infallible at a world but fallible at another will play a role.

notions are usually taken to have this property. But if possibility is a matter of sufficient similarity between cases, it does not: a possibility may be sufficiently similar to a second which is sufficiently similar to a third without the first being similar to the third. When the property holds, infallible methods are necessarily infallible; when it does not, a method can be both infallible and possibly fallible. As in standard epistemic logic, the property turns out to be crucial to derive knowledge of one's knowledge (section 5.4).[13]

We model possibility in the standard way by an accessibility relation over worlds. A method is infallible at a world if at all accessible worlds its unconditional outputs and its conclusions reached from true premises are all true:

Definition 3. Let W be the set of worlds, and $R \subseteq W \times W$ a reflexive accessibility relation over worlds. A method m is infallible at a world w iff:
(a) for any w', p such that wRw' and $p \in m(w')$, p is true at w'.[14]
(b) for any w', p, π such that wRw' and $p \in m(w', \pi)$, p is true at w' if all the propositions in π are true at w'.

Remark 2. A purely non-inferential method m is infallible iff (a) holds. A purely inferential method is infallible iff (b) holds.

Different constraints on relation R yield different notions of infallibility.

1. wRw' iff $w' = w$: a *true-belief-like* notion of knowledge, in which only the actual world needs to be considered. Though implausible as a philosophical account of knowledge, it is noteworthy that all our results can be obtained in that simple setting.[15]

2. wRw' for any w, w': a notion of knowledge that requires the metaphysical impossibility of error, and correspondingly precludes inductive knowledge. Arguably the notion defended by Descartes.

3. wRw' iff w' is "close" to w: a *safety* notion of knowledge such as the one defended by Williamson (2000) and Sosa (1996). If closeness is not transitive, what is possibly possible need not be possible, and knowledge of one's knowledge is not guaranteed (section 15).

[13] See Williamson (2000, ch.5) on the failures of epistemic introspection that result when the relevant notion of impossibility does not iterate.
[14] Recall that $m(w') = m(w', \emptyset)$ is the unconditional output of m at w'.
[15] Note that that option does not *equate* true belief and knowledge. Suppose that a method outputs both p and q, and that p is true but q false. Then one's belief that p on that basis is not knowledge, even though p is true. The option makes knowledge "true-belief-like", though, because it would classify many lucky true beliefs as "knowledge" just because no agent happens to use a given method in unfavourable circumstances.

4. In each context c, "knows" is associated with a specific R_c relation: a contextualist account along the lines of the one defended by DeRose (1995) and Lewis (1996).

Our results are independent of the choice. They only require that R be reflexive, except knowledge of one's knowledge which additionally requires transitivity.[16]

3 Applications: Gettier cases, fallible justification and inductive knowledge

We can already sketch how to use methods to model a few epistemological ideas.

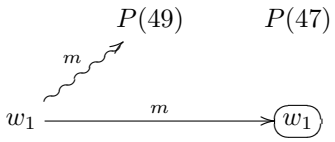
3.1 The prime number case and fake-barn-style Gettier cases

In our *Prime Numbers* case (sec. 1), Primo uses a method m that both produces a true belief that 47 is prime and a false belief that 49 is prime. (Let us assume

[16]Some related accounts of knowledge cannot be represented without substantial modifications of our apparatus. On Nozick's (1981, chap.3) view, one knows that p only if: *if p had been false, one would not have believed p*. On the Lewis-Stalnaker semantics of counterfactuals (Stalnaker, 1968; Lewis, 1973), the conditional is true if the corresponding material conditional $p \rightarrow q$ holds at each world up to the "closest" p world(s). This means that the range of possibility one has to look at for a given knowledge ascription *depends on the particular proposition at stake*, in our case, p. To model this, one needs to relativise accessibility to propositions: wR_pw' iff w' is at least as close to w than the first p world that is closest to w. Infallibility needs to be redefined: m is infallible at w *with respect to* p iff for any w', q, π, if wR_pw' and $q \in m(w', \pi)$ then q is true at w' if π is empty or all its members are true at w'. This invalidates our proof (below) that if a method is infallible at w, then the composition of Deduction and that method is infallible at w, since it may happen that a method is infallible *with respect to* p, without being infallible *with respect to a proposition q deduced from p*, if the first q-world is further away than the first p-world. That is why Deductive closure fails in Nozick's system.
Note than on the von Fintel-Gillies semantics of counterfactuals (von Fintel, 2001; Gillies, 2007), the condition is true iff $p \rightarrow q$ holds a set of close worlds fixed by context. The set does not depend on the particular p evaluated. In that setting Nozick's condition is modelled in our system by a context-relative R_c, and it does not violate closure.

Another view that is not straightforwardly accommodated by our models is the subject-sensitive or interest-relative accounts of (Hawthorne, 2004; Stanley, 2005). On that view the range of possibilities of error relevant to whether an agent knows p is affected by the stakes she has in p: the higher the cost of error, the broader set of possibilities of error becomes relevant her knowing. If stakes are relative to propositions (one's stake in p may be higher than one's stake in q), we get ranges of error that are p-dependent as in Nozick's semantics. So Infallibility must be redefined as before, and our proofs do not go through. If stakes are relative to a subject's situation, we can model stake-sensitivity with a stake relative accessibility relation $R_{s,t}$ akin to the contextualist one, but which is a function of subject and time instead of being function of context. On the latter model stake-relative versions of our results (notably Deductive closure) can be recovered.

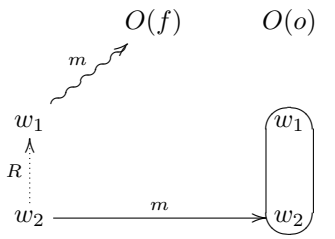
that he considered both questions at the actual world.) Let $P(47)$ and $P(49)$ be the relevant propositions:



(*Illustrations for methods models.* To avoid cluttering, I lay out the set of possible worlds W in a column that is repeated horizontally as many times as needed. Propositions are represented by circled sets of worlds; typically we use one column per proposition. Only the unconditional outputs of methods are represented: an arrow named m between w and p indicates that $p \in m(w)$, that is, $p \in m(w, \emptyset)$. When the output is true, the arrow is horizontal. Diagonal arrows indicate false beliefs and are signalled by wavy lines. When the output proposition is the empty set, as $P(49)$ is here, the arrow points directly to its name instead of a circle. When needed, accessibility relations between worlds are represented by dotted arrows labeled with R.)

At w_1, m outputs the belief $P(47)$ that is true at w_1, but it also outputs the belief $P(49)$, which is false. Consequently, m is fallible, and no belief based on m can be knowledge.

Similar models can be given for the fake-barn style of case (sec. 1). Suppose that at w_2 the agent looks at the real orange, but that there is an accessible world w_1 where she looks at a fake one and forms the false belief that it is an orange. Write $O(o)$ and $O(f)$ the relevant propositions:



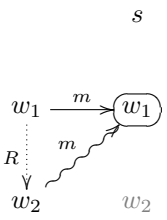
m produces only a true belief at w_2, but there is an accessible world in which it produces a false belief, namely w_1, and that is why the subject fails to know even in world w_2.

The models are straightforward but not trivial. In standard epistemic logic, Kp holds if and only if p is true at all accessible worlds. Unless impossible

worlds are introduced, it is true at every world that 47 is prime. So however accessibility is fixed, we get the result that it is known. Similarly, in Fitting's models for Logic of Proofs (Fitting, 2005, 4) , $t : p$ (which we can read as "t is the subject's justification for p") holds at a world iff t is evidence for p at w and p holds at all accessible worlds.[17] Here we get the result that it is know that 47 is prime as soon as some justification supports that belief, since the proposition that 47 is prime holds at every world. If we count Primo's calculations as justifications, we get the wrong results; if we don't, the models do not explain why they do not count as justifications.[18]

3.2 Standard Gettier cases

Consider (a slight variant of) Chisholm's (1966) sheep case: a man comes to believe that there is a sheep in a field by seeing a sheep-looking rock in the distance. As it happens, there is one, but it is hidden behind the rock. The case is straightforwardly modelled if we assume that the subject could have formed the same belief on the same basis in the absence of sheep. Let s be the proposition that there is a sheep in the field:



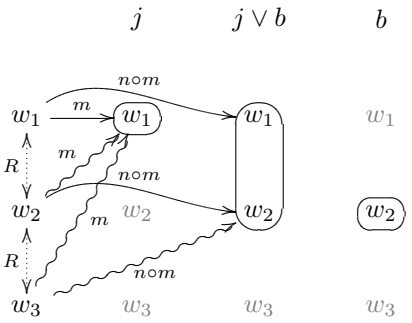
At w_2, the subject's method produces a false belief. At w_1, w_2 is an accessible possibility, so his method is fallible and he fails to know.

The original Gettier (1963) cases are modelled in basically the same way, except that we need to introduce a inferential step. Smith has good evidence

[17] The same holds for the extension of Fitting's models presented by Artemov and Nogina (2005, 1066).

[18] On Fitting's (Fitting, 2005, 5) strong models, any formula that is true at all accessible worlds has a justification, so Primo's belief would come out as justified. Artemov (2008, section 6) suggests that a knowledge-level justification for p is a "factive justification", i.e. "sufficient for an agent to conclude that p is true". This can be understood in three ways. (a) sufficient for an agent to *know* that p is true: the characterisation is circular. (b) such that necessarily, if p is justified by that justification, p is true: this wrongly ascribes knowledge in the *Prime Number* case. (c) such that *for any proposition* p', necessarily, if p' is justified by that justification, p' is true: this is infallibilism as we have defined it. Artemov's semantics (based on Fitting, 2005) suggests that he adopts the second construal.

that Jones owns a Ford, and infers that Jones owns a Ford (j) or Brown is in Barcelona (b). As it happens, b is true but j is false. Let m be the method that leads Smith to form the belief that Jones owns a Ford, and let n be the method that infers $j \vee b$ from j. Assuming that the subject could have formed the same beliefs while b was false:



At each world, by method m, the subject forms the belief that j, and by the method n applied to m, she infers $j \vee b$. In w_1 her evidence is not misleading: j is true. In w_2 her evidence is misleading, however $j \vee b$ is true because j is true. That is the Gettier situation. In w_3 the evidence is misleading as in w_2, but now b is not true, so $n \circ m$ outputs a false belief. Since w_3 is accessible from w_2, $n \circ m$ is not infallible at w_2, and that is why the subject fails to know.

Again the result is straightforward but not trivial. A common diagnosis of Gettier's original cases is the "no-false-lemma" view (or its generalisation the "no-false-assumption" view) according to which reasoning from false premises cannot provide knowledge (Clark, 1963).[19] Our models assumes nothing of the kind: what matters is whether the subject's *total inference* ($n \circ m$) could have lead to a false belief, not whether some intermediate steps are. Suppose I slightly overestimate heights, and from my belief that the door is over 2.5 meters high, I infer cautiously that it is at least more than 2m high: we may grant me knowledge of the latter even though the door was 2.4 meter high, for instance, and my initial belief false. (See Unger, 1968, 165 for a similar view.) Such verdicts are available in methods models.

[19] Artemov's formalisation of the cases in the context of the Logic of Proofs endorses a strong version of the no-false-lemma view: namely, that in the sense of justification relevant to knowledge there is no justification of a false proposition (Artemov, 2008, section 6). That threatens the possibility of inductive knowledge (see section 3.4). See Lycan (2006, section 6.1) for a recent defence of the no-false-assumption view.

3.3 Fallible but reliable methods

We have modelled Gettier cases in terms of fallible methods. That explains why they are not cases of knowledge, but not why they are not simply cases of unjustified belief, such as belief based on tea leaves readings.

There is a natural way to introduce the idea in our models, though we do not implement it formally here: each method gets assigned a *reliability measure* (a real between 0 and 1) at a world depending on its tendency to produce true beliefs rather than non-true ones at accessible worlds. In a finite setting, we could for instance take the measure to be the ratio of true beliefs to all beliefs produced at accessible worlds, but in general we need not assume that reliability is reducible to other notions except in very simple cases. For instance, if m is the method that leads one to believe that each ticket in a one-hundred ticket lottery is a looser, we may say that the reliability of m is .99. This gives a sense in which one's belief based on m that one's ticket will loose is justified without being knowledge.

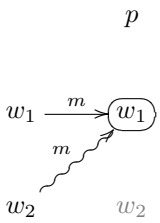
In section 5.1.2, we discuss how the notion of fallible justification, when plugged in a justified-true-belief account of knowledge, results in Gettier-type cases.

3.4 Inductive knowledge

Things being as they are, my mother often comes to *know* that the neighbours are home by seeing that their light is on. Absolutely speaking, it is of course physically possible that the neighbour's light is on and they are not there. But given their habits and the general circumstances, they could not be out with the light left on. This is what allows my mother to come to know that they are home simply by seeing their light. In the method infallibilist setting, that is cashed out as the idea that inductive knowledge is a matter of *local infallibility*, that is infallibility over a relevantly restricted set of accessible worlds.[20]

[20] What if the neighbours do have such a habit, but my mother has no idea of it and just rashly assumes that they are there? Then she does not know. On the method infallibilist view, that result can be obtained in two ways. Either we have an accessible world in which the neighbour's habits are different. Or we have an accessible world in which she believes something false on the basis of the same method, *e.g.*, that some *other* neighbours are there while they are not. Inductive knowledge would then require a method that is somehow sensitive to the neighbours' habits: for instance, that my mother would not draw the inference if they did not have the habit. Such a sensitivity may amount to knowledge of their habits, but it can easily fall short of *entailing* that they are there tonight if the light is on tonight. Even in these more realistic settings, inductive knowledge boils down to *local* infallibility, infallibility in the kind of world my mother is in. (The account of inductive knowledge assumed here

Let m be the method that produces my mother's belief that (p) the neighbours are there:



At w_1, the method produces a true belief. As in the model for Chisholm's sheep case (section 3.2), there is a world in which the method produces a false belief (w_2). But here that world is not a genuine possibility in w_1. The absence of neighbours while their light is on is not something that could have happened in the circumstances at hand. So m is infallible at w_1, and and p is known.

We can thus say that a method is *inductive* iff it is metaphysically fallible. If we want to call "strongly *a priori*" a piece of knowledge that is based on a non-inductive method, we get a vindication of the idea of contingent *a priori* knowledge: for instance, the method that would lead each subject to believe that she exists is metaphysically infallible.[21]

These considerations allow us to draw an important distinction between two properties that the label "fallibility" fails to distinguish. A reliable method can be "fallible" in the sense that it is (a) fallible properly, that is, it produces false beliefs *within the relevant set of worlds*, (b) inductive, that is, it produces false beliefs *outside* of the relevant set of worlds. An inductive method can thus be infallible in the proper sense. Fallibility in the first sense is incompatible with deductive closure and typically leads to Gettier cases; "fallibility" in the second sense (inductivity) is compatible with deductive closure and is better positioned to avoid Gettier cases.

is externalist: whether a subject has inductive knowledge by a certain method depends on features of her environment she may not be aware of. See Armstrong, 1973, 157, 166–7, 206–8; Dretske, 1971, 2–4.)

[21] I call metaphysically infallible methods *strongly a priori* to allow for the possibility of *weakly a priori* methods that would allow an agent to get knowledge without experience and yet depend on a certain kind of environment the agent is in. (Innate knowledge of grammar or physical properties of our environement may be a case in point.)

4 Methods models

All we have said so far assumed only a set of worlds, an alethic modality over them, and a set of propositions. That suffices to characterise methods, operations over them, their infallibility, and to give models of Gettier cases and inductive knowledge. To get a proper formal implementation of methods infallibilism, however, we now opt for a particular notion of proposition. What has been said so far nevertheless holds for other choices.

We take the simplest notion of proposition, namely a set of worlds. That has troublesome consequences with so-called Frege cases, though we point out that even there our simplest method models have something interesting to say. We show how agents are specified as sets of methods, and how knowledge and belief are from those. We give a language and we state basic equivalence results with Scott-Montague neighbourhood models (Montague, 1968, 1970; Scott, 1970).

4.1 Propositions and the problem of Frege cases

4.1.1 Propositions as sets of worlds

Definition 4. Propositions are sets of worlds: $P = \mathcal{P}(W)$.

A proposition p is true at a world w iff $w \in p$.

The resulting methods models are an extension of neighbourhood models. See Appendix B (section 8).

4.1.2 Frege cases

The choice keeps our models simple and on familiar grounds, but comes at a cost. Call *referential opacity cases* cases in which we are tempted to say that an agent knows (or believes) that p but fails to know (believe) that q, while p and q are true at exactly the same worlds.[22] Well-known candidates are:

Proper names Alice knows that Hespherus shines, but she does not know that Phosphorus shines. (Frege, 1892/1980)

Pierre believes that Londres is pretty, but he does not believe that London is pretty. (Kripke, 1979)

Indexicals David knows that that man's pants [unwittingly pointing at himself in the mirror] are on fire, but he does not know that his own pants are on fire. (Kaplan, 1989, sec XVII, see also Perry, 1979)

[22] The terminology comes from Quine (1953/1961).

Natural kinds terms Saul knows that there is water in the glass, but does not know that there is H_2O in the glass. (Kripke, 1980)

Logical equivalents Fred knows that p but does not know that $p \vee (\neg p \rightarrow p)$.

Given (Definition 4), if p and q are true at exactly the same world, $p = q$, and thus p is an output of a method m iff q is. So it appears that simple methods models cannot represent referential opacity cases.

The matter is not so straightforward, however. Suppose that, not knowing that Hespherus and Phosphorus are the same planet, Alice believes that Hespherus shines but not that Phosphorus shine. Let m^{H} be a constant for the method through which she believes that Hespherus shines (when she does), and m^{P} a constant for the method through which she believes that Phosphorus shines (when she does). (We introduce the full language in section 4.3.1.) In our language, we represent Alice's situation as follows:

$$\mathsf{Bm}^{\mathsf{H}}\mathsf{p} \wedge \neg\mathsf{Bm}^{\mathsf{H}}\neg\mathsf{p} \wedge \mathsf{Bm}^{\mathsf{P}}\neg\mathsf{p} \wedge \neg\mathsf{Bm}^{\mathsf{P}}\mathsf{p}$$

where p is a term for the proposition that Venus shines. In other terms: by method m^{H}, she believes the proposition to be true and does not believe it to be false; by method m^{P}, she believes it to be false and does not believe it to be true. That representation at least avoids a contradictory statement that $\mathsf{Bp} \wedge \neg\mathsf{Bp}$.[23] Moreover, it shows that the notion of method can at least partly capture Frege's elusive notion of "mode of presentation": a belief that p is a Hespherus-belief if based on m^{H}, and a Phosphorus-belief if based m^{P}. And it is compatible with a direct-referentialist view of content.

However, the suggested account has its limits. It is likely that referential opacity phenomena arise within a single method. For instance, if I see the end of a ship by one window, and its other end by another, I may rationally come to believe that they are two distinct ships (Perry, 1979, 483). I may thus believe that that ship [pointing at one end] is identical to that ship [pointing at the same end] while also believing that that ship [pointing at the other end] is not identical to that ship [poiting at the first end again]. We may want to count the two beliefs as being issued by the same method. So we get a case in which:

$$\mathsf{Bmp} \wedge \mathsf{Bm}\neg\mathsf{p}$$

[23] See Kripke (1979, section III).

where p is the proposition that that ship is identical to itself, and m the method by which I formed both beliefs. Here we cannot represent the contradiction as a contradiction of beliefs based on different sources. An option would be to make methods as fine-grained as necessary to make sure that Frege cases can all be cashed out in terms of distinct methods, but that strikes me as an *ad hoc* move, and would render the notion of method less natural.

Some further enrichment of the models thus appears needed to fully take into account referential opacity phenomena. One option is to distinguish propositions that are true at the same possible worlds: this can be done by introducing impossible worlds.[24] Another option is to take the outputs of methods to be sentence-like structures. Philosophically, that option corresponds equally to (a) a language-of-thought view, (b) a Frege-style view in which propositions are structured and not reducible to extensions, (c) a view of belief as a ternary relation between subjects, modes of presentations (represented by sentences), and coarse propositions.[25]

4.2 Frames and agents

Our frames are given by a set of worlds, an accessibility relation over them, and a set of methods for the agent. (In the multi-agent case, a set of methods can be introduced for each agent.) Propositions and the space of methods are themselves defined from the set of worlds:

Definition 5. A *methods frame* $\mathfrak{F}$ is a triple $\langle W, M^B, R\rangle$ where:

$M^B \subseteq \mathbf{M}$ is the *set of basic methods* of the agent, where $\mathbf{M} = W \times (\mathcal{P}(P) \times \mathcal{P}(P))$ with propositions as sets of worlds: $P = \mathcal{P}(W)$.[26]

$R \subseteq W \times W$ is a reflexive accessibility relation over worlds.

[24] We introduce a set of impossible worlds I and redefine the set of propositions as $P = \mathcal{P}(W \cup I)$. Premises and conclusions of methods are now taken from this extended set; however, they can remain functions from *possible* worlds. At impossible worlds, valuation is not compositional: any arbitrary set of formulas can hold — impossible worlds can be modelled as such sets. For two propositional constants p_1 and p_2 (standing for "Hespherus shines" and "Phosphorus shines", for instance) that hold at exactly the same possible worlds in a model, we have an impossible world w^* such that $w^* \in p_1$ and $w^* \notin p_2$, so that $p_1 \neq p_2$ and we can have $p_1 \in m(w)$ and $p_2 \notin m(w)$ for some method m and possible world w.

[25] We can take the premises and conclusions of methods to be *formulas* of our language, for instance — similar strategies appear in Awareness semantics (Fagin and Halpern, 1988) and Fitting's semantics for the Logic of Proofs (Fitting, 2005). (Though it should be noted that the technique is philosophically unsatisfactory, since it makes things look as though the mental states of a subject were dependent on the language of the ascriber.) However, the resulting semantics is prone to self-referential paradoxes, and methods may have to be typed in order to avoid them.

[26] Cf. sections 2.1 and 4.1.

A *transitive methods frame* is a methods frame in which R is transitive.

The primitives of our models are just worlds, a set of basic methods for each agent, and a background alethic modality given by $\langle W, R\rangle$ that will be used to characterise infallibility. The rest is derived as follows.

Definition 6. The agent's *method set* is $M = M^{B\circ+}$.[27]

An agent is represented by a set of methods. We assume an agent free of modular or computational limitations: the set of her methods is closed under composition and union. As we noted in section 2.1, bounded agents can be modelled by putting restrictions on building M out of M^B, and more sophisticated representations of agents could have their method set changing through time.

We define a set of infallible methods at each world:

Definition 7. $M^I(w)$ is the set of *infallible methods at* w:
$m \in M^I(w)$ iff for all w' such that wRw':
(a) $\forall p(p \in m(w') \rightarrow w' \in p)$, and
(b) $\forall p, \pi((p \in m(w', \pi) \wedge w' \in \bigcap \pi) \rightarrow w' \in p)$.

Corollary 1. *Infallible methods preserve infallibility. If* $m \in M^I(w)$ *and* $n \in M^I(w)$, $m + n \in M^I(w)$ *and* $m \circ n \in M^I(w)$.

Proof. Obvious from Definitions 2 and 7 for method union. For method combination, let m, n, w be such that $m \in M^I(w)$ and $n \in M^I(w)$. We show that $m \circ n \in M^I(w)$:

(a) Suppose wRw' and $p \in m \circ n(w')$ for some p, w'. By Definition 2, there is a π such that $\pi = n(w')$ and $p \in m(w', \pi)$. By Definition 7 (a), since $n \in M^I(w)$, wRw' and $\pi = n(w')$, we have $w' \in \bigcap \pi$. By Definition 7 (b), since $m \in M^I(w)$, wRw', $w' \in \bigcap \pi$ and $p \in m(w', \pi)$, we have $w' \in p$. Generalising over p: $\forall p(p \in m \circ n(w') \rightarrow w' \in p)$.

(b) Suppose wRw', $p \in m \circ n(w', \pi)$ and $w' \in \bigcap \pi$. By Definition 2, there is a π' such that $\pi' = n(w', \pi)$ and $p \in m(w', \pi')$. By Definition 7 (b), since $n \in M^I(w)$, wRw', $w' \in \bigcap \pi$ and $\pi' = n(w', \pi)$, $w' \in \bigcap \pi'$. By Definition 7 (b) again, since $m \in M^I(w)$, wRw', $w' \in \bigcap \pi'$ and $p \in m(w', \pi')$, $w' \in p$. Generalising over p, π: $\forall p, \pi((p \in m \circ n(w', \pi) \wedge w' \in \bigcap \pi) \rightarrow w' \in p)$. □

Lemma 1. *If* $\mathfrak{F}$ *is a transitive methods frame, then for any* m, w, w' *such that* wRw', *if* $m \in M^I(w)$ *then for any* w' *such that* wRw', $m \in M^I(w')$.

[27] cf. section 2.2.

Proof. Let m, w be such that $m \in M^I(w)$, and w' such that wRw'. Let w'' be any world such that $w'Rw''$. By the transitivity of R, wRw''. Since $m \in M^I(w)$, by Defintion 7, $\forall p(p \in m(w'') \rightarrow w'' \in p)$ and $\forall p, \pi((p \in m(w'', \pi) \wedge w'' \in \bigcap \pi) \rightarrow w'' \in p)$. By Definition 7 again, $m \in M^I(w')$. □

Lemma 1 is crucial to the derivation of knowledge of one's knowledge (section 18).

Finally we define a range of functions from worlds and methods to sets of propositions for belief and knowledge on a basis, and from worlds to sets of propositions for belief and knowledge *simpliciter*. "B" and "K" each refer to functions of both types, but the ambiguity is convenient and their arguments always disambiguate:

Definition 8. $B : (m, w) \mapsto \{p : m \in M \wedge p \in m(w)\}$ gives the *agent's beliefs on the basis of* m.

$K : (m, w) \mapsto \{p : p \in B(m, w) \wedge m \in M^I(w)\}$ gives the *agent's knowledge on the basis of* m.

$B : w \mapsto \{p : \exists m(p \in B(m, w))\}$ gives the *agent's beliefs simpliciter*.

$K : w \mapsto \{p : \exists m(p \in K(m, w)\}$ gives the *agent's knowledge simpliciter*.

Beliefs are just the outputs of the agent's methods on the basis of no premises. Knowledge is belief on an infallible basis. $B(m, w)$ is the set of propositions believed on the basis of m at w, and $B(w)$ is the set of propositions believed at w, and similarly for knowledge. $B(w)$ and $K(w)$, as well as the functions $w \mapsto B(m, w)$ and $w \mapsto K(m, w)$ for a given m, are neighbourhood functions: see Appendix B (section 8) .

The definitions of belief and knowledge *simpliciter* are not entirely innocuous. For all we have said, an agent may be such that one of her methods unconditionally outputs p and another one $\neg p$. The definition implies that such an agent both believes p and not-p. Some may want to resist that; one may want to say for instance that an agent believes p iff one of her methods outputs p and no other outputs $\neg p$. However, our existentially quantified definition is by far the simplest; alternative options create holistic constraints on belief and knowledge that would prevent us to get fully general theorems such as $\mathsf{Kp} \rightarrow \mathsf{KKp}$. However, I do not see the fact that we rely on this choice as a important liability, though; for one lesson of methods models is that deep generalisations about knowledge should be stated in terms of *based* belief and knowledge rather than in terms of belief and knowledge *simpliciter*.

4.3 Language

4.3.1 Syntax

Definition 9. Let $\mathsf{M} = \{\mathsf{m}, \mathsf{n}, \ldots\}$ be a set of methods constants. The set of methods terms is given by the grammar:

$$\mu ::= \mathsf{m} | \mu + \nu | \mu \circ \nu$$

$\mathcal{L}$ is the set of formulas given by:

$$\phi ::= \mathsf{p} | \top | \neg\phi | \phi \vee \psi | \phi \wedge \psi | \phi \rightarrow \psi | \phi \leftrightarrow \psi | \mathsf{B}\mu : \psi | \mathsf{K}\mu : \phi | \mathsf{B}\psi | \mathsf{K}\phi | \Box\phi$$

where $\mathsf{P} = \{\mathsf{p}, \mathsf{q}, \mathsf{r}, \ldots\}$ is a set of propositional constants.

We introduce two B and K operators, one for methods-relative belief and knowledge and the other for belief and knowledge *simpliciter*. $\Box$ will express the background alethic modality.

By convention, $\mathsf{B}\mu :$ and $\mathsf{K}\mu :$ take the narrowest scope: we read $\mathsf{B}\mu : \phi \rightarrow \psi$ as $(\mathsf{B}\mu : \phi) \rightarrow \psi$ and not as $\mathsf{B}\mu : (\phi \rightarrow \psi)$.

4.3.2 Semantics

Definition 10. Let $\mathfrak{M} =< \mathfrak{F}, V >$ be a model where $V : \mathsf{P} \cup \mathsf{M} \rightarrow P \cup \mathbf{M}$ is a valuation function that assigns a proposition to each propositional constant and a method to each method constant. We define $[\![\cdot]\!]^{\mathfrak{M}}$:

Methods terms. $[\![\mathsf{m}]\!]^{\mathfrak{M}} = V(\mathsf{m})$,
$[\![\mu + \nu]\!]^{\mathfrak{M}} = [\![\mu]\!]^{\mathfrak{M}} + [\![\nu]\!]^{\mathfrak{M}}$,
$[\![\mu \circ \nu]\!]^{\mathfrak{M}} = [\![\mu]\!]^{\mathfrak{M}} \circ [\![\nu]\!]^{\mathfrak{M}}$.
Propositional logic. $[\![\mathsf{p}]\!]^{\mathfrak{M}} = V(\mathsf{p})$, $[\![\top]\!]^{\mathfrak{M}} = W$,
$[\![\neg\phi]\!]^{\mathfrak{M}} = W \backslash [\![\phi]\!]^{\mathfrak{M}}$,
$[\![\phi \vee \psi]\!]^{\mathfrak{M}} = [\![\phi]\!]^{\mathfrak{M}} \cup [\![\psi]\!]^{\mathfrak{M}}$,
and as usual for other logical connectives.
Necessity, belief and knowledge.
$[\![\Box\phi]\!]^{\mathfrak{M}} = \{w : \forall w'(wRw' \rightarrow w' \in [\![\phi]\!]^{\mathfrak{M}})\}$,
$[\![\mathsf{B}\mu : \phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in B([\![\mu]\!]^{\mathfrak{M}}, w)\}$,
$[\![\mathsf{K}\mu : \phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in K([\![\mu]\!]^{\mathfrak{M}}, w)\}$,
$[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in B(w)\}$,
$[\![\mathsf{K}\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in K(w)\}$.

Truth. $\models^{\mathfrak{M}}_{w} \phi$ iff $w \in [\![\phi]\!]^{\mathfrak{M}}$.
Validity. $\models^{\mathfrak{M}} \phi$ iff for any w, $\models^{\mathfrak{M}}_{w} \phi$.

The clause for $\Box\phi$ is familiar from Kripke models. The clauses for $\mathsf{B}\phi$ and $\mathsf{K}\phi$ are familiar from neighbourhood models, and it is easy to see that the clauses for $\mathsf{B}\mu : \phi$ and $\mathsf{K}\mu : \phi$ pick up the neighbourhood function corresponding to the unconditional output of the method m designated by μ, namely: $w \mapsto m(w)$, provided that m is one of the agent's methods (belief case) and that it is infallible (knowledge case).

In Appendix B (section 8) we compare neighbourhood models and methods models in more detail. We show that for any neighbourhood model for B, there is an equivalent method model for B, and conversely, and that for any neighbourhood model for K in which $\mathsf{K}\phi \rightarrow \phi$ is valid, there is an equivalent method model for K, and conversely (Theorems 20 and 21). But we also argue that methods models are more explanatory than neighbourhood ones, because they allow us to derive the modal axioms from the structure of an agent's methods.

We check that belief and knowledge on a basis entail belief and knowledge *simpliciter*:

Corollary 2. *For any* $\mathfrak{M}, w$, $\models^{\mathfrak{M}}_{w} \mathsf{B}\mu : \phi \rightarrow \mathsf{B}\phi$ *and* $\models^{\mathfrak{M}}_{w} \mathsf{K}\mu : \phi \rightarrow \mathsf{K}\phi$ *for any* μ,ϕ.

Proof. From Definition 10, if $\models^{\mathfrak{M}}_{w} \mathsf{B}\mu : \phi$ then there is a $m \in M$ such that $[\![\phi]\!]^{\mathfrak{M}} \in m(w)$. By Definition 8, $[\![\phi]\!]^{\mathfrak{M}} \in B(w)$, and by Definition 10 again, $\models^{\mathfrak{M}}_{w} \mathsf{B}\phi$. And analogously for K. □

The methods algebra of Appendix A (section 7) and the semantics give us a series of equivalences:

Corollary 3. *The following are valid in any method model* $\mathfrak{M}$*:*
$\mathsf{B}(\mu + \nu) + \rho : \phi \leftrightarrow \mathsf{B}\mu + (\nu + \rho) : \phi$,
$\mathsf{B}\mu + \nu : \phi \leftrightarrow \mathsf{B}\nu + \mu : \phi$,
$\mathsf{B}\mu + \mu : \phi \leftrightarrow \mathsf{B}\mu : \phi$,
$\mathsf{B}(\mu \circ \nu) \circ \rho : \phi \leftrightarrow \mathsf{B}\mu \circ (\nu \circ \rho) : \phi$,
$\mathsf{B}(\mu + \nu) \circ \rho \leftrightarrow \mathsf{B}(\mu \circ \rho) + (\nu \circ \rho) : \phi$,
and similarly for K*, for any* μ, ν, ρ, ϕ.
For some methods model $\mathfrak{M}$, $\not\models^{\mathfrak{M}} \mathsf{B}\mu \circ (\nu + \rho) : \phi \leftrightarrow \mathsf{B}(\mu \circ \nu) + (\mu \circ \rho) : \phi$, *and similarly for* K*, for some* μ, ν, ρ, ϕ.

5 Results

The consequences of methods models fall within four groups, corresponding to different idealisations of agents:

1. For *any agent*: knowledge entails belief and truth (*subjectivity* and *factivity* of knowledge). That is a welcome result, since these are the only two (quasi-)uncontroversial facts about knowledge. Moreover, we have *referential transparency*: if p and q are true exactly at the same worlds, p is known iff q is. That is a limitation of the simpler models, as we noted (section 4.1).

2. For *perfect reasoners*, who have specific methods to believe all logical truths and all the logical consequences of what they believe: *deductive closure*. With unbounded resources, this validates the logical omniscience axiom **K**: $\mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$.

3. For *perfect introspecters* and *perfect confident introspecters*, who have methods to ensure that they believe that they believe p whenever they believe p (positive psychological introspection), that they believe that they *do not* believe p when they do not (negative psychological introspection), that they believe that they *know* p whenever they believe p (positive confident introspection) and that they believe that they do not know p when they do not believe it (negative confident introspection): *self-knowledge* ($\mathsf{B}\phi \leftrightarrow \mathsf{KB}\phi$), *partial negative epistemic introspection* ($\neg\mathsf{B}\phi \to \mathsf{K}\neg\mathsf{K}\phi$) and, if possible possibilities are possible, *epistemic positive introspection* or axiom **4** ($\mathsf{K}\phi \to \mathsf{KK}\phi$).

4. For *excellent agents*, whose methods are all infallible: *believing is knowing* ($\mathsf{B}\phi \leftrightarrow \mathsf{K}\phi$) and *epistemic negative introspection* or axiom **5** ($\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$).

With *pure reasoners*, we get a normal modal logic **K** for the belief *simpliciter* operator and **KT** for the knowledge *simpliciter* operator, and putting all idealisations together, we get a standard **S5** system for both. This gives us an equivalence between those models and standard Hintikka models for the subpart of $\mathcal{L}$ that is free of method terms, and shows that methods models can be as powerful as the standard ones.

As the summary indicates, the idealisations which we use to derive our results are natural idealisations of the psychology of an agent. (For positive epistemic

introspection, we need also an assumption on the structure of possibilities.) They are more intuitive than direct judgement on the **S5** axioms or on formal constraints on accessibility relations (transitivity, euclideanity). Correspondingly, they give us a better understanding of why and when the various axioms of standard epistemic logic hold.[28]

5.1 Subjectivity, factivity, and referential transparency

5.1.1 Referential transparency: the rule of equivalence

Theorem 1. *Referential transparency (**E**$_{\mathsf{BK}}$). If* $\models^{\mathfrak{M}} \phi \leftrightarrow \psi$*, then* $\models^{\mathfrak{M}} \mathsf{B}\phi \leftrightarrow \mathsf{B}\psi$ *and* $\models^{\mathfrak{M}} \mathsf{K}\psi \leftrightarrow \mathsf{K}\phi$*, for any methods model* $\mathfrak{M}$*.*

Proof. Suppose $\models^{\mathfrak{M}} \phi \leftrightarrow \psi$. We have $[\![\phi]\!]^{\mathfrak{M}} = [\![\psi]\!]^{\mathfrak{M}}$, so $[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in B(w)\} = \{w : [\![\psi]\!]^{\mathfrak{M}} \in B(w)\} = [\![\mathsf{B}\psi]\!]^{\mathfrak{M}}$, and similarly for K. □

Referential transparency is the rule of equivalence of classical modal logic. It is known to determine the class of all neighbourhood frames.[29] Thus by Theorem 20 the schema determines methods frames for the B operator.

Referential Transparency is a consequence of our choice of modelling propositions as sets of possible worlds. It ensures that our models are classical and similar to neighbourhood models. However, in doxastic and epistemic terms, that means that belief and knowledge are *referentially transparent* in the sense discussed above. This is a problematic consequence, as we said (section 4.1).

5.1.2 Factivity and subjectivity

Theorem 2. *Subjectivity (**S**).* $\models^{\mathfrak{M}} \mathsf{K}\phi \rightarrow \mathsf{B}\phi$ *for any methods model* $\mathfrak{M}$*.*

Proof. Evident from Definitions 8 and 10. □

The theorem states that knowledge is subjective, in the sense that knowledge is in part a matter of the agent's psychology, namely, his beliefs.[30]

[28] The properties of the epistemic accessibility relation in standard epistemic logic can be naturally interpreted as *indistinguishability* relations. But this hides an important ambiguity. Indistinguishability can be understood as inability to know the difference: w is indistinguishable from w' to one iff in w one cannot know that w is different from w'. Or indistinguishability can be understood as sameness of internal state: w is indistinguishable from w' to one iff in w one is in the same internal state as in w'. The first reading is Hintikka's (2007) and the second is roughly Lewis' (1996). Each has problematic consequences, as we noted in section 1. (Thanks to an anonymous referee here.)

[29] Chellas (1980, 257).

[30] By saying that knowledge is "subjective" I only mean that it requires a state of mind — not that it is "subjective" in the sense in which matters of taste are said to be so. Subjectivity

Theorem 3. *Factivity ($\boldsymbol{T}_{\mathsf{K}}$).* $\models^{\mathfrak{M}} \mathsf{K}\phi \rightarrow \phi$ *for any methods model* $\mathfrak{M}$.

Proof. Evident from the fact that R is reflexive and Definitions 7, 8 and 10. □

Factivity entails that knowledge is consistent, or axiom **D**: $\mathsf{K}\phi \rightarrow \neg\mathsf{K}\neg\phi$. (Assume $\mathsf{K}\phi$; by factivity, ϕ, so $\neg\neg\phi$, and by factivity again, $\neg\mathsf{K}\phi$.)

It is known from neighbourhood semantics that $\mathbf{E}_{\mathsf{BK}}$ and $\mathbf{T}_{\mathsf{K}}$ determine truthful neighbourhood models.[31] By Theorem 21 the schemas determine methods models with respect to the K operator.

Note that we need two things to derive Factivity. First we need the reflexivity of R, that is, □ should be a modality that itself satisfies **T** :

Theorem 4. *Alethic necessity ($\boldsymbol{T}_{\Box}$).* $\models^{\mathfrak{M}} \Box\phi \rightarrow \phi$ *for any method model* $\mathfrak{M}$.

Proof. From the reflexivity of R and Definition 10. □

This reflects the fact that □ is intended as an alethic modality. Second, we need *strict infallibility*, i.e. that the methods in $M^I(w)$ are such that *all* their unconditional outputs are true.

Suppose we try to put a weaker condition on knowledge; for instance, we include in $M^I(w)$ all the methods that are highly reliable at w (see section 3.3). Factivity can no longer be derived: $p \in m(w)$ and $m \in M^I(w)$ do not entail that p is true at w, since some of the unconditional outputs of m may be false at w. Truth must then be added as a *separate* condition on knowledge: $p \in K(w)$ iff there is a m such that $p \in m(w)$, $m \in M^I(w)$ *and* $w \in p$. This is in essence the "justified-true-belief" analysis of knowledge, and it is open to Gettier counterexamples because its two conjuncts (m is a reliable/adequate method, and p is true) can be simultaneously satisfied by coincidence.

5.1.3 Failure of logical omniscience and introspection

None of the other axioms of modal logic are valid.

Theorem 5. *Each of* ***M***, ***C***, ***K***, ***N***, ***4***, ***5*** *for* B *and* K *fails in some methods model* $\mathfrak{M}$, *and* ***D*** *and* ***T*** *for* B *fail in some methods model* $\mathfrak{M}$.

is in contrast with the notion of "implicit knowledge" that Hintikka models are often taken to formalise, which does not require than an agent be in any sense aware of the things she "knows".

[31] A neighbourhood model $\langle W, N\rangle$ is truthful iff $w \in \bigcap N(w)$ for any w. See Appendix B (section 8).

$\boldsymbol{D}_{\mathsf{B}}$. $\mathsf{B}\phi \to \neg\mathsf{B}\neg\phi$

$\boldsymbol{T}_{\mathsf{B}}$. $\mathsf{B}\phi \to \phi$

$\boldsymbol{M}_{\mathsf{B}}$. $\mathsf{B}(\phi \wedge \psi) \to (\mathsf{B}\phi \wedge \mathsf{B}\psi)$	$\boldsymbol{M}_{\mathsf{K}}$. $\mathsf{K}(\phi \wedge \psi) \to (\mathsf{K}\phi \wedge \mathsf{K}\psi)$
$\boldsymbol{C}_{\mathsf{B}}$. $(\mathsf{B}\phi \wedge \mathsf{B}\psi) \to \mathsf{B}(\phi \wedge \psi)$	$\boldsymbol{C}_{\mathsf{K}}$. $(\mathsf{K}\phi \wedge \mathsf{K}\psi) \to \mathsf{K}(\phi \wedge \psi)$
$\boldsymbol{K}_{\mathsf{B}}$. $\mathsf{B}(\phi \to \psi) \to (\mathsf{B}\phi \to \mathsf{B}\psi)$	$\boldsymbol{K}_{\mathsf{K}}$. $\mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$
$\boldsymbol{N}_{\mathsf{B}}$. $\mathsf{B}\top$	$\boldsymbol{N}_{\mathsf{K}}$. $\mathsf{K}\top$
$\boldsymbol{4}_{\mathsf{B}}$. $\mathsf{B}\phi \to \mathsf{B}\mathsf{B}\phi$	$\boldsymbol{4}_{\mathsf{K}}$. $\mathsf{K}\phi \to \mathsf{K}\mathsf{K}\phi$
$\boldsymbol{5}_{\mathsf{B}}$. $\neg\mathsf{B}\phi \to \mathsf{B}\neg\mathsf{B}\phi$	$\boldsymbol{5}_{\mathsf{K}}$. $\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$

Proof. Neighbourhood models that invalidate each schemas are known. By Theorem 20 they can be used to show that none of the B schema are valid in methods models. Moreover, it is easy to construct such counter-models as truthful neighbourhood models, so that by theorem 21 we have counterexamples to the K schemas. □

In appendix C (sec. 9), we give illustrations of violations of **M**, **N**, **K**, and **4**. In particular, the violation of **K** illustrates our *Watson* case (sec. 1).

5.2 Perfect reasoning

The first interesting class of methods models is that of *perfect reasoners*. Intuitively, an agent is a perfect reasoner if she believes all logical truths, and deduces all logical consequences of what she believes. To model such agents, we define the two following methods:

Definition 11. The *Pure Reason* method m^R is the method such that $m^R(w, \pi) = \{W\}$ for any w, π.[32]

For any model $\mathfrak{M}$, we write m^{R} the constant such that $[\![\mathsf{m}^{\mathsf{R}}]\!]^{\mathfrak{M}} = m^R$.

That is, Pure Reason outputs the tautology given any set of premises, including the empty set.

Definition 12. The *Multi-Premise Deduction* method m^D is the method such that $m^D(w, \pi) = \{p : \exists q, r \in \pi(q \cap r \subseteq p)\}$ for any w, π.

For any model $\mathfrak{M}$, we write m^{D} the constant such that $[\![\mathsf{m}^{\mathsf{D}}]\!]^{\mathfrak{M}} = m^D$.

[32] Thus defined, Pure Reason entails that the agent exists at any world. To avoid this, one could instead use a Conditional Pure Reason method: $m^{CR}(w, \pi) = \{W\}$ for any $w, \pi \neq \emptyset$. Conditional Pure Reason outputs the tautology only if the agent has some other belief. The resulting schema are $\models^{\mathfrak{M}} \mathsf{B}\phi \to \mathsf{B}\top$ and $\models^{\mathfrak{M}} \mathsf{B}\phi \to \mathsf{K}\top$ for any ϕ.

That is, Deduction maps a set of premises to all logical consequences of any pair of premises. (Note that it outputs nothing on the basis of the empty set of premises.)[33]

We call *perfect reasoner model* a methods model such that $m^R, m^D \in M$.[34]

Theorem 6. *Knowledge of logic ($\boldsymbol{N}_{\mathsf{B}}$ and $\boldsymbol{N}_{\mathsf{K}}$). For any perfect reasoner model $\mathfrak{M}$, $\models^{\mathfrak{M}} \mathsf{Bm}^{\mathsf{R}} : \top$ and $\models^{\mathfrak{M}} \mathsf{Km}^{\mathsf{R}} : \top$. By Corollary 2, $\models^{\mathfrak{M}} \mathsf{B}\top$ and $\models^{\mathfrak{M}} \mathsf{K}\top$.*

Proof. We first prove that at any world, the agent believes the tautology on the basis of Pure Reason; we then prove that Pure Reason is infallible at any world.

Let w be any world in a perfect reasoner model $\mathfrak{M}$. By Definition 11, $W \in m^R(w)$. By Definition 8, $W \in B(w)$. By Definition 10, $\models^{\mathfrak{M}}_w \mathsf{B}\top$.

Furthermore, by Definition 11, for any p, w, π, if $p \in m^R(w, \pi)$ then $p = W$, so $w \in p$. Thus by Definition 7, $m^R \in M^I(w)$. By Definition 8, $W \in K(w)$, and by Definition 10, $\models^{\mathfrak{M}}_w \mathsf{K}\top$. □

Theorem 7. *Deductive closure. For any perfect reasoner model $\mathfrak{M}$:*

$$\models^{\mathfrak{M}} \mathsf{B}\mu : (\phi \to \psi) \to (\mathsf{B}\nu : \phi \to \mathsf{Bm}^{\mathsf{D}} \circ (\mu + \nu) : \psi)$$

$$\models^{\mathfrak{M}} \mathsf{K}\mu : (\phi \to \psi) \to (\mathsf{K}\nu : \phi \to \mathsf{Km}^{\mathsf{D}} \circ (\mu + \nu) : \psi)$$

for any μ, ν, ϕ, ψ. Thus by Corollary 2, $\models^{\mathfrak{M}} \mathsf{B}(\phi \to \psi) \to (\mathsf{B}\phi \to \mathsf{B}\psi)$ and $\models^{\mathfrak{M}} \mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$ for any ϕ, ψ.

Proof. We show that whenever $p \to q$ and p are believed on the basis of m and n respectively, q is believed out of $m^D \circ (m+n)$. We additionally show that m^D is infallible at any world, which entails that $m^D \circ (m+n)$ is infallible whenever m and n are (Corollary 1).

Let $\mathfrak{M}, w$ be a perfect reasoner model and a world such that $\models^{\mathfrak{M}}_w \mathsf{B}\mu : (\phi \to \psi) \wedge \mathsf{B}\nu : \phi$ for some μ, ν, ϕ, ψ. By Definition 10 and Definition 8, there are m, n

[33] In particular, given any premise, Deduction will output the tautology. But that does not make Pure Reason redundant. If all purely non-inferential methods of the agent are fallible, then composing Deduction with them cannot yield knowledge, since the resulting composed method is fallible as well. Adding Pure Reason to the method set of such an agent enables knowledge of tautologies. (Thanks to an anonymous referee here.)

[34] Note that an agent may be a perfect reasoner without having Pure Reason and Deduction in its *basic* set M^B. To illustrate, let $m, n \in M^B$ be such that, for some $p \subseteq W$, $m(w, \pi) = \{W\}$ if $w \in p$ and $\emptyset$ otherwise, and $n(w, \pi) = \{W\}$ if $w \notin p$ and $\emptyset$ otherwise. (m outputs the tautology at p worlds, n outputs the tautology at not-p worlds.) We have $m + n = m^R$. So if M is the union and composition closure of M^B, $m + n \in M$, and the corresponding model is a perfect reasoner model provided that $m^D \in M$ as well, even if $m^R \notin M^B$.

and p, q such that $[\![\mu]\!]^{\mathfrak{M}} = m$, $[\![\nu]\!]^{\mathfrak{M}} = n$, $[\![\phi]\!]^{\mathfrak{M}} = p$, $[\![\psi]\!]^{\mathfrak{M}} = q$, $(W \backslash p) \cup q \in m(w)$ and $p \in n(w)$ and $m, n \in M$. By Definition 2, $(W \backslash p) \cup q, p \in (m+n)(w)$. Since $((W \backslash p) \cup q) \cap p \subseteq q$, by Definition 12 we have $q \in m^D \circ (m+n)(w)$. By Definition 10, $\models^{\mathfrak{M}}_w \mathsf{Bm}^{\mathsf{D}} \circ (\mu + \nu) : \psi$, which completes the proof of $\mathbf{K}_{\mathsf{B}}$.

Since m^D is purely inferential, we need only establish its inferential infallibilty (see Definition 3). Let $\mathfrak{M}, w$ be any perfect reasoner model and world. Suppose that $p \in m^D(w', \pi)$, wRw' and $w' \in \bigcap \pi$ for some p, π, w'. By Definition 12, there are $q, r \in \pi$ such that $p \supseteq q \cap r$. Since $w' \in \bigcap \pi$, $w' \in p$. Generalising over p, π, w', by Definition 3, $m^D \in M^I(w)$.

Now suppose $\models^{\mathfrak{M}}_w \mathsf{K}\mu : (\phi \rightarrow \psi) \wedge \mathsf{K}\nu : \phi$ for a perfect reasoner model $\mathfrak{M}$, a world w, and some ϕ,ψ. The situation is as before with, additionally, $m, n \in M^I(w)$ (Definitions 10 and 8). Since $m^D \in M^I(w)$ and infallible methods preserve infallibility (Corollary 1), $m^D \circ (m+n) \in M^I(w)$. We show as before that $m^D \circ (m+n) \in M$ and $q \in (m^D \circ (m+n))(w)$, so by Definition 10 $\models^{\mathfrak{M}}_w \mathsf{Km}^{\mathsf{D}} \circ (\mu + \nu) : \psi$, which completes the proof of $\mathbf{K}_{\mathsf{K}}$. □

The proof of $\mathbf{K}_{\mathsf{K}}$ relies on two things. First, for any methods m, n, the agent has a union method $m+n$ that outputs the union of the original outputs of m and n. This means that the agent "puts together" the result of any two methods. (Note that this does *not* mean that she believes the conjunction of original outputs.) Second, given any method m, $m^D \circ m$ outputs all the logical consequents of any two conclusions reached by m. We show that $m^D \circ m$ is infallible if m is, and that if p, q are unconditional outputs of m, then $m^D \circ m$ outputs all logical consequents of $p \wedge q$.

It is easy to see that $m^D \circ m$ will output the consequents of any *two* premises given my m, $m^D \circ m^D \circ m$ the consequents of any *three* premises, $m^D \circ m^D \circ m^D \circ m$ the consequents of any four premises, and so on. Correspondingly, we can limit the number of consequences the agent is able to reach by putting limits on the number of repeated applications of Deduction she can make, and we can model a dynamic process of reasoning by indexing those limits to time.

Methods models thus allow us to draw a distinction between *deductive closure proper* and *logical omniscience*. If an agent is limited on the number of m^D steps that she can reach, she will not be logically omniscient. But still, any consequence of what she knows that she *does* deduce will be knowledge, since m^D preserves infallibility. The idea that *any* agent knows all the consequences she does deduce can be somewhat indirectly captured by the following theorem:

Theorem 8. *Deduction preserves knowledge. For any methods model* $\mathfrak{M}$*:*

$$\models^{\mathfrak{M}} \mathsf{K}\mu : \phi \rightarrow (\mathsf{Bm}^{\mathsf{D}} \circ \mu : \psi \rightarrow \mathsf{Km}^{\mathsf{D}} \circ \mu : \psi)$$

(If anything is known on the basis of μ, then anything believed on the basis of deduction from μ is known).

Proof. Let $\mathfrak{M}, w$ be such that $\models^{\mathfrak{M}}_{w} \mathsf{K}\mu : \phi$ for some ϕ. By Definitions 10 and 8, $[\![\mu]\!]^{\mathfrak{M}} \in M^I(w)$. Since Deduction is infallible (Theorem 7) by Corollary 1, we have $m^D \circ [\![\mu]\!]^{\mathfrak{M}} \in M^I(w)$. Now suppose that $\models^{\mathfrak{M}}_{w} \mathsf{Bm}^{\mathsf{D}} \circ \mu : \psi$ for some ψ. Since $[\![\mathsf{m}^{\mathsf{D}} \circ \mu]\!]^{\mathfrak{M}} = m^D \circ [\![\mu]\!]^{\mathfrak{M}} \in M^I(w)$, $\models^{\mathfrak{M}}_{w} \mathsf{Km}^{\mathsf{D}} \circ \mu : \psi$. □

The theorem has no equivalent in a language without methods terms. (The closest we can formulate is $\mathsf{K}\phi \rightarrow ((\Box(\phi \rightarrow \psi) \wedge \mathsf{B}\psi) \rightarrow \mathsf{K}\psi)$, which is counterexampled if the agent believes ψ from some other reasons than ϕ, as in our *Watson* case. Note that $\mathsf{K}\phi \rightarrow (((\phi \rightarrow \psi) \wedge \mathsf{B}\psi) \rightarrow \mathsf{K}\psi)$ reduces to $\mathsf{K}\phi \rightarrow ((\psi \wedge \mathsf{B}\psi) \rightarrow \mathsf{K}\psi)$, which, barring bizarre cases, holds only for excellent agents: see Definition 20 (p. 41).

Further exploration of the deductive aspects of methods models are made in Appendix D (section 10):

1. Axioms **M** and **C** for belief and knowledge follow from axiom **K**. But it is also possible to get them separately, by defining a Single-Premise Deduction method m^{SD} (for **M**) and a Conjunctive Deduction method m^{CD} (for **C**). The relation between m^D, m^{SD} and m^{CD} is: $m^D = m^{SD} \circ m^{CD}$ (Corollary 4).

2. These results are correlated to topological properties of sets of sets, as in neighbourhood semantics.

3. The B and K versions of the **K** axiom are mutually independent, and similarly for **M**, **C**, **N**.

4. Having Pure Reason and Deduction is *sufficient* to satisfy the **N** and **K** axioms, respectively, but not *necessary*. So we have not characterised the full class of methods models that validates the axioms. However, we argue that that is not a defect: an agent might well satisfy **N** and **K** without having specific methods for doing so, but that would then be a sort of coincidence. The important regularities about knowledge are the methods-relative ones, not the ones stated in terms of belief and knowledge *simpliciter*.

5.3 Consistency

Pure Reason and Deduction do not guarantee than an agent is consistent: that is, the $\mathbf{D}_B$ axiom for belief ($\mathsf{B}\phi \rightarrow \neg\mathsf{B}\neg\phi$) may fail. (The $\mathbf{D}_K$ axiom for knowledge is of course guaranteed by the factivity of knowledge, Theorem 3.) Can we define a method to validate axiom $\mathbf{D}_B$? No. And that is an intuitive result: avoiding contradictions among one's beliefs is not a matter of forming one's beliefs, but rather of *revising* them. As long as we have not defined *methods for belief revision* — for instance, functions from a method set to another —, we cannot define a method that ensures consistency. We can at most give a (trivial) constraint to satisfy consistency:

Definition 13. A *consistent agent model* is a methods model such that for w, $\bigcap B(w) \neq \emptyset$.

Theorem 9. *If $\mathfrak{M}$ is a consistent agent model, ($\boldsymbol{D}_B$) $\models^{\mathfrak{F}} \mathsf{B}\phi \rightarrow \neg\mathsf{B}\neg\phi$ for any ϕ.*

Proof. Suppose $\mathfrak{M}$ is a consistent agent model and w a world such that $\models^{\mathfrak{M}}_{w} \mathsf{B}\phi$ for some ϕ. By Definition 10 there is a p such that $[\![\phi]\!]^{\mathfrak{M}} = p$ and $p \in B(w)$. By Definition 13 it follows that $(W \backslash p) \notin B(w)$, and by Definition 10 $\not\models^{\mathfrak{M}}_{w} \mathsf{B}\neg\phi$. □

5.4 Perfect introspection and perfect confidence

The second interesting classes of models are that of *perfect introspecters* and *confident introspecters*. Intuitively, an agent is a perfect introspecter if whenever she believes something, she believes that she does, and whenever she does not believe something, she believes that she does not. An agent is a confident introspecter if whenever she believes something, she believes that she knows it, and whenever she does not believe something, she believes that she does not know it.

Some abbreviations will be useful:

Definition 14. $b_p := \{w : p \in B(w)\}$,
$-b_p := \{w : p \notin B(w)\} = W \backslash b_p$,
$k_p := \{w : p \in K(w)\}$,
$-k_p := \{w : p \notin K(w)\} = W \backslash k_p$,

b_p is the proposition that p is believed, $-b_p$ its negation, and analogously for k_p and $-k_p$.

5.4.1 Perfect introspection: self-knowledge

Definition 15. Given any methods model $\mathfrak{M}$, we define:

For each method m, the *Positive Introspection of* m, noted $pi(\mathfrak{M}, m)$, is the method such that for any w, p, π: $p \in pi(\mathfrak{M}, m)(w, \pi)$ iff for some p', $p = b_{p'}$ and $p' \in m(w)$.

For each set of methods X, the *Negative Introspection of* X, noted $ni(X)$, is the method such that for any w, p, π: $p \in ni(\mathfrak{M}, X)(w, \pi)$ iff for some p', $p = -b_{p'}$ and there is no $m \in X$ such that $p' \in m(w)$.

Convention. We write $pi(m)$ and $ni(X)$ for $pi(\mathfrak{M}, m)$ and $ni(\mathfrak{M}, m)$ when the intended model is clear from the context.

For each method m, the method $pi(m)$ outputs that the agent believes p whenever m outputs p. In intuitive terms, an agent that has $pi(m)$ takes herself to have method m, as far as her beliefs are concerned — whether she does in fact have m or not. For each set of methods X, the method $ni(X)$ outputs that the agent does not believe p, when p is not among the outputs of the methods in X. In intuitive terms, an agent that has $ni(X)$ takes herself to have at most the methods in X.

It is easy to see that $pi(m)$ is infallible if m is one of the agent's methods and $ni(X)$ is infallible if X includes all of the agent methods:

Lemma 2. *For any methods model and any world w:*

For any $m \in M$, the positive introspection of m is infallible: $pi(m) \in M^I(w)$.

For any $X \supseteq M$, the negative introspection of X is infallible: $ni(x) \in M^I(w)$

Proof. Positive and Negative Introspection methods are purely non-inferential, so we need only prove their non-inferential infallibility.

Positive introspection. Let $\mathfrak{M}, m$ be such that $m \in M$ and let w be any world. Suppose that wRw' and $p \in pi(m)(w')$ for some w', p. By Definition 15, there is some p' such that $p = b_{p'}$ and $p' \in m(w')$. Since $m \in M$ and $p' \in m(w')$, by Definition 8, $p' \in B(w')$. Hence $w' \in b_{p'}$. Generalising over w', p, $pi(m) \in M^I(w)$ (Definition 7).

Negative introspection. Let $\mathfrak{M}, X$ be such that $M \subseteq X$ and let w be any world. Suppose that wRw' and $p \in ni(X)(w')$ for some w', p. By Definition 15, there is a p' such that $p = -b_{p'}$ and for no $m \in X$, $p' \in m(w')$. Since $M \subseteq X$, for no $m \in M$, $p' \in m(w')$, so by Definition 8, $p' \notin B(w')$. Hence $w' \in -b_{p'}$. Generalising over w', p, $ni(X) \in M^I(w)$ (Definition 7). □

An agent is a *perfect introspecter* if she has positive and negative introspection methods that perfectly match her set of methods. Perfect introspecters have perfect knowledge of their own beliefs:

Definition 16. A *perfect introspecter model* is a methods model such that for each $m \in M$, $pi(m) \in M$, and $ni(M) \in M$.

Theorem 10. *Self-knowledge (**SK**). If $\mathfrak{M}$ is a perfect introspecter model, $\models^{\mathfrak{M}}_w$ $\mathsf{B}\phi \rightarrow \mathsf{KB}\phi$ and $\models^{\mathfrak{M}}_w \neg\mathsf{B}\phi \rightarrow \mathsf{K}\neg\mathsf{B}\phi$.*

Proof. Positive self-knowledge. Suppose $\mathfrak{M}$ is a perfect introspecter model and w any world such that $\models^{\mathfrak{M}}_w \mathsf{B}\phi$. By Definitions 10 and 8, there are p, m such that $[\![\phi]\!]^{\mathfrak{M}} = p$, $m \in M$ and $p \in m(w)$. By Definition 15, $b_p \in pi(m)$, where $b_p = [\![\mathsf{B}\phi]\!]^{\mathfrak{M}}$ (Definitions 8, 10 and 14) and by Definition 16 $pi(m) \in M$. Since by Lemma 2, $pi(m) \in M^I(w)$, and since $[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} \in pi(m)(w)$ and $pi(m) \in M$, we have $\models^{\mathfrak{M}}_w \mathsf{KB}\phi$ (Definitions 8 and 10).

Negative self-knowledge. Suppose $\mathfrak{M}$ is a perfect introspecter model and w any world such that $\models^{\mathfrak{M}}_w \neg\mathsf{B}\phi$. By Definitions 10 and 8, there are no p, m such that $[\![\phi]\!]^{\mathfrak{M}} = p$, $m \in M$ and $p \in m(w)$. By Definition 15, $-b_p \in ni(M)$, where $-b_p = [\![\neg\mathsf{B}\phi]\!]^{\mathfrak{M}}$ (Definitions 8, 10 and 14) and by Definition 16, $ni(M) \in M$. Since by Lemma 2, $ni(M) \in M^I(w)$, and since $[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} \in ni(M)(w)$ and $ni(M) \in M$, $\models^{\mathfrak{M}}_w \mathsf{K}\neg\mathsf{B}\phi$ (Definitions 8 and 10). □

A few remarks on introspection methods are in order.

Introspection methods are characterized in *agent*- and *model*-relative terms. Let p, m, w be such that $p \in m(w)$. At w, the Positive Introspection of m will output the proposition that the agent believes p, namely b_p. But since we model propositions as sets of worlds, b_p is the set of worlds in which the agent believes p. Which set that is depends on what model we are considering. Since the output of an introspection method cannot be defined independently of a given model, these methods cannot be so defined either. In multi-agent settings, their output would be further relativized to agents, and b_p and $pi(), ni()$ should be parametrized accordingly. So we cannot have model- and agent-independent method constants for Introspection methods, by contrast with Deduction and Pure Reason. This reflects the fact that Introspection methods tell us something about the agents.[35]

[35] A contrast may be useful. Let us write $m : p$ for the proposition that m unconditionally outputs p. Given any method m, we can define the 'method-introspection' (as opposed to belief-introspection) method $!m$ such that $!m$ outputs $m : p$ whenever m outputs p. That is,

For the same reason, though $pi()$ and $ni()$ can be considered as operations on methods and sets of methods, they are not *model-independent* operations such as Union and Composition.

Relatedly, it is not straightforward to *construct* perfect introspection models. Take a model $\mathfrak{M}$ built on the basic method set M^B and a method such that m is a method of the agent but the agent lacks the corresponding Introspection method: $m \in M$ but $pi(\mathfrak{M}, m) \notin M$. To ensure that the agent introspects her own method m, we cannot simply take the model $\mathfrak{M}'$ built on the basic set $M^B \cup \{pi(\mathfrak{M}, m)\}$. For it may be that $pi(\mathfrak{M}, m)$ is *not* an Introspection method in the new model: $pi(\mathfrak{M}, m) \neq pi(\mathfrak{M}', m)$. And analogously for Negative Introspection. Thus there is no straightforward way to close a model under Introspection.

We have used *fine-grained* Positive Introspection methods: one per method. We could have used instead coarser methods, along the lines of Negative Introspection: $pi^*(X)$ is such that $p \in pi^*(X)(w, \pi)$ iff $p' \in m(w)$ for some $m \in X$ and p' s.th. $p = b_{p'}$. The method $pi^*(X)$ outputs that the agent believes p whenever *any* method within X outputs p. This is sufficient for knowledge of one's beliefs. But finer-grained methods are required for knowledge of one's knowledge, as we will see.

By contrast, (perfect) Negative Introspection is essentially holistic. We cannot get perfect negative introspection on a method-per-method basis. Suppose we define, for a given method m, a method that outputs that the agent does not believe p whenever *the method* m does not output p. The method will go wrong in cases in which some *other* method of the agent outputs p, so that the agent believes p after all. At most, we can define a fine-grained infallible method that outputs that the agent does not believe p *on the basis of* m.[36] But negative introspection methods of this type would not deliver the proposition that the agent does not believe p *simpliciter*.[37]

The asymmetry between Positive and Negative Introspection on this score reflects the fact that in order for an agent to *believe* p, it is sufficient that *some*

$!m$ tells us *that* m *outputs* p whenever m does output p. (! is analogous to the proof-checker in the Logic of Proofs (Artemov, 1994).) The ! operator can be defined in a model-independent way. But ! is not a operator of *psychological* introspection: its outputs are about what *methods* output, not about what *agents* believe, and it 'introspects' any method, irrespective of whether the agent has it or not.

[36] Such methods would be analogues to the negative verifier ? in the Logic of Proofs (Fitting, 2008).

[37] More precisely, given our extensional notion of proposition, they will deliver this proposition only if it happens to be coextensive with the proposition that the agent does not believe p on the basis of some particular method m.

of her method outputs p, while for her *not* to believe p, it is necessary that *none* of her methods ouptuts p.

5.4.2 Perfectly Confident Introspection: knowledge of one's knowledge and partial knowledge of one's ignorance

Confident Introspection methods are introspection methods whose output is the proposition that one knows (or fails to know) instead of the proposition that one believes (or fails to believe):

Definition 17. Given any methods model $\mathfrak{M}$, we define:

For each method m, the *Positive Confident Introspection of* m, noted $pc(\mathfrak{M}, m)$, is the method such that for any w, p, π, $p \in pc(\mathfrak{M}, m)(w, \pi)$ iff for some p', $p = k_{p'}$ and $p' \in m(w)$.

For each set of methods X, the *Negative Confident Introspection of* X, noted $nc(X)$, is the method such that $p \in nc(\mathfrak{M}, X)(w, \pi)$ iff for some p', $p = -k_{p'}$ and there is no $m \in X$ such that $p' \in m(w)$.

Convention. We omit reference from the model when it is clear from the context, and write $pc(m)$ and $nc(X)$ for $pc(\mathfrak{M}, m)$ and $nc(\mathfrak{M}, m)$, respectively.

In intuitive terms, an agent that has Negative Introspection of X believes as if all its knowledge came from methods in X. Negative Confidence is infallible applied to any X that includes the agent's methods:

Lemma 3. *For any methods model and any world w:*

For any $X \supseteq M$, the Negative Confident Introspection of X is infallible: $nc(x) \in M^I(w)$.

Proof. We transpose the proof of Lemma 2, using the fact that if $p' \notin B(w')$ then $p' \notin K(w')$ (Definition 8 and Theorem 2) . □

For Positive Confidence we would expect a *conditional* infallibility result: if m is infallible, then $pc(m)$ is infallible. This is guaranteed, however, only if the frame is *transitive*. For any output of m, $pc(m)$ tells that the agent knows it. Thus if m is infallible, all the outputs of $pc(m)$ are true. But for $pc(m)$ to be *infallible*, its outputs must be true at all accessible worlds; so m has to be infallible at all accessible worlds. If the frame is transitive, the infallibility of m at a world ensures its infallibily at accessible worlds (Lemma 1), so we get the desired result:

Lemma 4. *For any transitive methods models and any world w:*

For any $m \in M$, if $m \in M^I(w)$ then the Positive Confident Introspection of m is infallible: $pc(m) \in M^I(w)$.

Proof. Positive Confidence methods are purely inferential, so we need only prove the infallibility of their unconditional outputs.

Let $\mathfrak{M}, m$ be a model in a transitive frame and a method such that $m \in M$ and let w be a world such that $m \in M^I(w)$. Suppose that wRw' and $p \in pc(m)(w')$ for some w', p. By Definition 15, there is some p' such that $p = k_{p'}$ and $p' \in m(w')$. Since $\mathfrak{M}$ is transitive, $m \in M^I(w)$ and wRw', $m \in M^I(w')$ (Lemma 1). Since $m \in M$, $p' \in m(w')$ and $m \in M^I(w')$, by Definition 8, $p' \in K(w')$. Hence $w' \in k_{p'}$. Generalising over w', p, $pc(m) \in M^I(w)$ (Definition 7). □

An agent is a Perfect Confident Introspecter if she has Confident Introspection methods that exactly match her method set:

Definition 18. A *Perfect Confident Introspecter model* (or Confident Introspecter, for short) is a methods model such that for each $m \in M$, $pc(m) \in M$, and $nc(M) \in M$.

Theorem 11. *Confident Introspection. If $\mathfrak{M}$ is a confident introspecter model, $\models^{\mathfrak{M}} \mathsf{B}\phi \rightarrow \mathsf{BK}\phi$ and $\models^{\mathfrak{M}} \neg\mathsf{B}\phi \rightarrow \mathsf{B}\neg\mathsf{K}\phi$.*

Proof. Evident from Definitions 10, 8 and 18. □

Theorem 12. *Knowledge of one's knowledge (4). If $\mathfrak{M}$ is a Confident Introspecter model in a transitive frame, $\models^{\mathfrak{M}} \mathsf{K}\phi \rightarrow \mathsf{KK}\phi$ for any ϕ.*

Proof. Suppose $\mathfrak{M}$ is a perfect confident introspecter model in a transitive frame and w any world such that $\models^{\mathfrak{M}}_{w} \mathsf{K}\phi$. By Definitions 10 and 8, there are p, m such that $[\![\phi]\!]^{\mathfrak{M}} = p$, $m \in M$, $p \in m(w)$ and $m \in M^I(w)$. By Definition 17, $k_p \in pc(m)$, where $k_p = [\![\mathsf{K}\phi]\!]^{\mathfrak{M}}$ (Definitions 8, 10 and 14) and by Definition 18, $pc(m) \in M$. Since the frame is transitive, $m \in M$ and $m \in M^I(w)$, by Lemma 4, $pc(m) \in M^I(w)$. Since $[\![\mathsf{K}\phi]\!]^{\mathfrak{M}} \in pc(m)(w)$ and $pc(m) \in M$, $\models^{\mathfrak{M}}_{w} \mathsf{KK}\phi$ (Definitions 8 and 10). □

The result is the only one that requires a stronger assumption on the background accessibility relation than reflexivity. The proof relies on finer-grained introspection methods. An indiscriminate confidence method ($pc^*(M)$ such that

$p \in pc^*(M)(w,\pi)$ iff $p' \in m(w)$ for some $m \in X$ and p' s.th. $p = k_{p'}$) would be fallible as long as some of the agent's methods are fallible.[38]

Theorem 13. *Partial knowledge of one's ignorance (**p5**). If $\mathfrak{M}$ is a Confident Introspecter, $\models^{\mathfrak{F}} \neg \mathsf{B}\phi \rightarrow \mathsf{K}\neg\mathsf{K}\phi$ for any ϕ.*

Given subjectivity ($\mathsf{K}\phi \rightarrow \mathsf{B}\phi$), the theorem is equivalent to a conditionalised version of axiom **5**: $\models^{\mathfrak{F}} \neg \mathsf{B}\phi \rightarrow (\neg\mathsf{K}\phi \rightarrow \mathsf{K}\neg\mathsf{K}\phi)$. I am using "ignorance" here in a slightly unnatural way to refer to everything the subject fails to know. (Thus if p is false p is part of the subject's "ignorance" in that sense.)

Proof. We transpose the proof of negative self-knowledge (Theorem 10), using Lemma 3. □

Partial knowledge of one's ignorance (**p5**) is a very intuitive result. There has been much debate around axiom **5** of epistemic logic, according to which if one does not know p, one knows that one does not to know it. The intuition that has lead many to think that it was appropriate for knowledge is, I think, the following: ask an agent whether p, she will "look up" her memory to see whether it contains p, and if it does not, she will answer (rightly) that she does not know. But that is precisely the idea that our result cashes out: *when a subject fails to know p because they fail even to believe it*, they know that they do not know p.

A few remarks on Confidence methods.

The Confident Introspection methods we define are introspection methods, in the sense that the methods $pc(m)$ and $nc(X)$ "track" what method m and the methods in X are doing. It would be more natural to define them as a composition of a psychological introspection method and a confidence method: whenever the agent believes p, she believes that she does (psychological introspection), and whenever she believes that she believes p, she infers that she knows p (confidence).

That can be done for negative confidence: we could define an inferential method nc^* such that $nc^*(w,\pi) = \{-k_p : -b_p \in \pi\}$. The method is infallible: suppose that $p \in nc^*(w,\pi)$ and all the premisses in π are true. Then there is

[38] At a given world w, knowledge of one's knowledge holds if the agent is indiscriminately confident of all her methods that happen to be infallible at w: that is, the agent has $pc^*(M \cap M^I(w))$. But since the agent's method set is not world-dependent in our models, this would mean that at *every world* the agent is confident of her methods that are infallible at w. Furthermore, to derive knowledge of one's knowledge at every world, the agent's method set should include all methods of this type: for any w, $pc^*(M \cap M^I(w)) \in M$. It is not clear to me whether such an idealisation makes sense.

some p', p'' such that $p' \in \pi$, $p' = -b_{p''}$, and $p = -k_{p'}$. Since p' is true, $-b_{p'}$ is true. Since $-b_{p'}$ is true, $-k_{p'}$ is true. Thus the combined method $nc^* \circ ni(M)$ is infallible, which gives us (**p5**) for Perfect Introspecters with nc^*. In addition, we get the schema $\models^{\mathfrak{M}} \mathsf{B}\neg\mathsf{B}\phi \rightarrow \mathsf{B}\neg\mathsf{K}\phi$ for any agent with nc^*. This makes explicit how Confident Introspection is parasitic on Introspection.

Unfortunately, the parallel idea for Positive Confidence cannot be implemented in our models. This is a consequence of the coarse individuation of proposition as sets of worlds (sec. 4.1). Consider the method pc^* such that $pc^*(w, \pi) = \{k_p : b_p \in \pi\}$. Suppose that we have a methods model $\mathfrak{M}$ with two methods $m, n \in M$ such that only m ever outputs p, only n ever outputs q, but they do so at the very same worlds. We have $b_p = b_q$: the proposition that the agent believes that p and the proposition that she believes that q are the same, even if p and q are different. Now at a world where m outputs p, $pi(m)$ outputs b_p, and therefore $pc^* \circ pi(m)$ outputs k_p. But since $b_p = b_q$, $pc^* \circ pi(m)$ will also output k_q. As a result, if m is infallible but n is not, k_p will be true but k_q will be false. So $pc^* \circ pi(m)$ is not guaranteed to be infallible if m is.

Roughly put, what is going on is that the difference between introspecting from m ($pi(m)$) and introspecting from n ($pi(n)$) is "lost" on the confidence method pc^* when their outputs are not differentiable ($b_p = b_q$). That is why we had to build our Confidence methods as directly introspecting the lower-order methods, i.e. as Confident Introspection methods.

The idea that Confidence is parasitic on psychological Introspection can still be partly captured through the following constraint on method models:

Definition 19. A *Normal Confidence* model is a method model such that for any m, X, $pc(m) \in M$ only if $pi(m) \in M$ and $nc(X) \in M$ only if $nc(X) \in M$.

A natural class of normal confidence models is that of Perfect Introspecters that are Confident Introspecters. Other ways of cashing out the dependence of Confidence upon Introspection require more refined models in which we will not get into here.[39]

[39] One is to have *fine-grained propositions*, for instance formulas. Roughly, in our example we would have $pi(m)(w) = \{\mathsf{Bp}\}$ and $pi(m)(w) = \{\mathsf{Bq}\}$ at some world w, where $\mathsf{Bp} \neq \mathsf{Bq}$ even though they hold at the same worlds. Positive Confidence can be defined as follows: $pc^{**}(w, \pi) = \{\mathsf{Kp} : \mathsf{Bp} \in \pi\}$ and knowledge of one's knowledge is easily derived.

5.5 Excellence

A third interesting class of models is that of *excellent agents*. An excellent agent is simply an agent whose methods are all infallible.

Definition 20. A *Excellent Agent model* is a methods model such that $M \subseteq M^I(w)$ for any w.

Theorem 14. *Belief is knowledge (**BK**). If $\mathfrak{M}$ is an excellent agent model, $\models^{\mathfrak{M}} \mathsf{B}\phi \leftrightarrow \mathsf{K}\phi$.*

Proof. The right-to-left direction follows from Subjectivity.

The left-to-right direction is evident from Definition 20, 8 and 10. Let $\mathfrak{M}, w$ be an Excellent Agent model and world such that $\models^{\mathfrak{M}}_w \mathsf{B}\phi$ for some ϕ. There is a p such that $[\![\phi]\!]^{\mathfrak{M}} = p$ and $p \in B(w)$. Since $p \in B(w)$, there is a $m \in M$ such that $p \in m(w)$. By Excellence, $m \in M^I(w)$. So $p \in K(w)$, and $\models^{\mathfrak{M}}_w \mathsf{K}\phi$. □

For an excellent agent, believing is knowing, since all her beliefs are infallibly based. Correlatively, the only way such an agent fails to know something is by *failing to believing it* — while imperfect agents can fail to know something by having a false belief or by having a fallibly-based belief in it. That is the basis of the next result:

Theorem 15. *Perfect knowledge of one's ignorance (**5**). If $\mathfrak{M}$ is an Excellent Confident Introspecter model, $\models^{\mathfrak{M}} \mathsf{K}\phi \rightarrow \mathsf{K}\neg\mathsf{K}\phi$.*

Proof. Suppose $\mathfrak{M}, w$ are an Excellent Confident Introspecter model and a world such that $\models^{\mathfrak{M}}_w \neg\mathsf{K}\phi$. By Excellence and Theorem 14, $\models^{\mathfrak{M}}_w \neg\mathsf{B}\phi$. By Confident Introspection and Theorem 13, $\models^{\mathfrak{M}}_w \mathsf{K}\neg\mathsf{K}\phi$. □

The result is again intuitive. However, it provides an illuminating perspective over the much-disputed axiom **5**. While the axiom is assumed in many successful applications of epistemic logic, it faces a glaring and simple counterexample: false belief. If I mistakenly believe that my car keys are in my pocket, then I do not *know* that they are there, but (typically at least) I will not know that I do not know it. To the contrary, I (typically) think that I do know it. That does not reflect any irrationality on my part; nor is it plausible to say that I implicitly know that I do not know that they are there. Our result is in line with that idea: we derive knowledge of one's ignorance for excellent agents, i.e. agents who *cannot have false beliefs*. At the same time, the result explains why

(and when) it is safe to assume **5**: namely, when the agent can be taken to be excellent with respect to the relevant facts. For instance, in many game-theoretic applications, it is assumed that the agents cannot have false beliefs about the game setup nor draw false inferences. The simple **S5** epistemic system is suited to that use.

5.6 Discussion

We have shown that under a natural set of idealisations, the axioms of a standard **S5** epistemic logic hold. That ensures that many applications of epistemic logic can be recovered in methods models. Additionally, we have derived a number of principles linking knowledge and belief: Subjectivity, Self-knowledge, Partial Knowledge of One's Ignorance and Belief is Knowledge.

But even though their stronger versions are equivalent to a single-operator **S5** system, methods models provide an insight into the idealisations at work behind the **S5** axioms. A widespread picture about epistemic logic is the following:

> The axioms of standard epistemic logic represent an ideal of rationality, where rationality is a matter of internal or subjective coherence and unbouded computational ability, as opposed to a matter of excellence, that is, objective success.

Our result suggest a radically different picture.

We have three sets of derivations that are independent: (a) Perfect Reasoning (**K** and **N**), (b) Positive epistemic introspection and partial negative epistemic introspection (**4** and **p5**), (c) Excellence (**BK**). Negative epistemic introspection (**5**) is obtained from a composition of (b) and (c). But it is important to note that the results in (b) and (c) do *not* assume perfect reasoning, nor does (c) assume introspection or confidence. We thus have three distinct sets of idealisations at play.

Moreover, it can be shown that while the Perfect Reasoning methods are *non-informative*, Introspection and Confident Introspection are *informative*, in the sense that they "narrow down" the sets of possible worlds compatible with what the agent believes and knows. (That is, if a possible world is incompatible with a agent's belief based on $m^D \circ m$, that possible world is also incompatible some agent's belief based on m; the same does not hold for $pi(m)$ and $pc(m)$.) See Appendix E (section 11) for a formal definition of the relevant notions of

information and informativeness.

Together, these remarks suggest the following picture:

1. Deduction and Reason (axiom **K**) are properly a matter of pure rationality or internal coherence. They do not provide information, but make explicit the information the subject has.

2. Introspection and Confident Introspection (axioms **4** and **p5**) are a matter of excellence with respect to the inner. Both assume that the agent has reliable ways to find out about its own internal states. Both are information-purveying methods. Thus satisfying axiom **4** or **KK** is not simply a matter of internal coherence.

3. Excellence (**BK** and with Confident Introspection, axiom **5**) typically requires excellence with respect to the outer. Extreme cases aside, it requires one's methods to provide information.[40] It is not a simple matter of internal coherence either.

The methods approach thus exhibits a distinction between three groups of idealisations behind standard epistemic logic: pure rationality, internal excellence and external excellence. It can be used as a guide as to when the axioms are appropriately assumed to hold.

6 Conclusion

Methods models provide a formal representation of knowledge that rests on the methods-infallibilist conception of knowledge. The conception is in line with various trends in mainstream epistemology, such as reliablism, safety theories or some variants of virtue epistemology. I have argued that it is suited to model and think about classical epistemological issues such as the Gettier problem or inductive knowledge. Because the intuitive notion of method, or basis of belief, takes a centre stage in the models, they should prove more amenable to epistemologists than the standard Hintikka models. However, I have also shown that standard epistemic logic systems can be recovered from methods models through a series of natural idealisations of agents and a constraint on possibility. The models thus offer a new vindication of the standard axioms and an illuminating perspective on why and when they hold or not. They should consequently prove useful to formal epistemologists as well.

[40] The extreme case is that of an agent who has only Reason and Deduction.

The models can be further developed in a range of directions, and much needs yet to be done. On the formal side, they should be studied syntactically, starting from the algebra we sketched and by introducing an operator to express infallibility. Soundness and completeness properties should be established. Relatedly, the models may be usable as models for the Logic of Proofs. A further important formal development is to build variants of the models that integrate common treatments of referential opacity, which will most likely require us to leave the ground of neighbourhood semantics.

On the epistemology side, four developments can be mentioned. First, our methods are only methods of belief *formation*. Methods of belief *inhibition* and *revision* should also be considered. The former would allow holistic constraints on the belief system (*e.g.*, if a method produces a belief that p and another a belief that $\neg p$, both beliefs are suspended); but for that reason, they may be hard to accommodate formally. The latter would require to recast our models in dynamic terms, with temporal slices of agents being characterised by the set of beliefs reached at each point or by distinct sets of methods. Second, we have only considered *fine-grained* Introspection and *maximal* Confidence. Variants of Introspection that fail to discriminate between beliefs produced by a range of similar methods should be discussed, as well as cautious agents whose epistemic confidence extends only to the beliefs produced by a subset of their methods. Third, our methods can straightforwardly be used to model *conditional* belief and hypothetical reasoning. If an agent has a method m such that $p \in m(w, \pi)$, then (at w) she conditionally believes that p on the hypothesis that π. However, the Introspection methods as defined here are unsuitable for that purpose (they imply that, for any p, the agent conditionally believes that she believes that p on the hypothesis that p). Accordingly, we may want to redefine them as a class of non-inferential methods. Fourth, the idea of a reliability measure over methods that we have sketched in section 3.3 should be investigated in order to see whether it can yield an interesting notion of epistemic probability.

These developments only concern the single-agent case. A further one is of course to study multi-agent settings and to characterise common knowledge in method terms. *Mind-reading* methods may prove relevant here, as well as *perspectives* (section 2.1).

Finally, the methods approach need not be restricted to epistemology. Along methods of belief formation, one may characterise an agent by a set of *methods for decision*, whose inputs are a set of premises (and perhaps a set of aims) and whose outputs are actions. Truth is here replaced by success. In the epis-

temological case, the approach induces a shift of focus from individual beliefs to classes of beliefs formed in the same way. In the practical case, we get an analogous shift of focus from particular intentions or actions to classes of actions that result from a same policy. The latter kind of focus is already familiar from rule utilitarianism and virtue theories. Methods models may provide useful representations of such ideas.

7 Appendix A. Algebra for methods

$\langle \mathbf{M}, +, \circ, 0, 1 \rangle$ is an algebraic structure over the set of methods, where 0 and 1 are defined as follows: [41]

Definition 21. The *empty method*, noted 0, is the method such that $0(w, \pi) = \emptyset$ for all w, π.

The *identity method*, noted 1, is the method such that $1(w, \pi) = \pi$ for all w, π.

Here are the main properties of the algebra.

Theorem 16. *Method union is idempotent, commutative and associative.*

For any $m, n, r \in \mathbf{M}$: $m + m = m$, $m + n = n + m$, *and* $(m + n) + r = m + (n + r)$.

Proof. From the corresponding properties of set union and Definition 2. □

Remark 3. The *empty method* 0 is uniquely characterised as the method such that $0 + n = n$ for any n.

Theorem 17. *Method composition is associative but not idempotent nor commutative.*

For any $m, n, r \in \mathbf{M}$: $m \circ (n \circ r) = (m \circ n) \circ r$. *But* $m \circ m = m$ *and* $m \circ n = n \circ m$ *are not valid.*

Proof. Associativity: from the associativity of function composition and Definition 2.

Counterexample to idempotence: for any proposition $p \in P$, write $\neg p$ the negation of p. Let m be such that for any w, π, $m(w, \pi) = \{\neg p : p \in \pi\}$. At any w we have: $m(w, \{p\}) = \{\neg p\} \neq (m \circ m)(w, \{p\}) = \{\neg\neg p\}$.[42]

Counterexample to commutativity: consider m defined as above, and n such that at any w, $n(w, \pi) = \{p \wedge q : p, q \in \pi\}$ where $p \wedge q$ denotes the conjunction of any propositions p and q. Assuming $p \neq q$, we have $(m \circ n)(w, \{p, q\}) = \{\neg(p \wedge q), \neg p, \neg q\} \neq (n \circ m)(w, \{p, q\}) = \{\neg p \wedge \neg q, \neg p, \neg q\}$ at any w. □

Remark 4. The *identity method* 1 is uniquely characterised as the method such that $1 \circ n = n \circ 1 = n$ for any $n \in \mathbf{M}$.

[41] Thanks to Paul Egré and Johan van Benthem for suggesting this development.

[42] The counterexamples given in this section assume a few uncontroversial facts about propositions, such as: the negation of a proposition is a proposition and at least some negation of a proposition is distinct from its own negation. These will hold however propositions are fleshed out.

Remark 5. Purely non-inferential methods are methods who are insensitive to what premises they are given. They can thus be characterised as the set of methods m such that $m \circ n = m$ for any $n \in \mathbf{M}$.

For any $n \in \mathbf{M}, w, \pi$: $0 \circ n = 0$ (the empty method is purely non-inferential) and $(n \circ 0)(w, \pi) = n(w, \emptyset)$ (applying a method n to the empty one amounts to applying n without premise).

Theorem 18. *Method union does not distribute over method composition.*
$m + (n \circ r) = (m + n) \circ (m + r)$ *is not valid.*

Proof. Take $r = 1$; the claim reduces to $m + n = (m + n) \circ (m + 1)$, which is guaranteed only if $m + n$ is non-inferential. □

Theorem 19. *Composition distributes right-to-left over union, but not left-to-right.*
For any $m, n, r \in \mathbf{M}$: $(m+n) \circ r = (m \circ r) + (n \circ r)$. *By contrast,* $m \circ (n+r) = (m \circ n) + (m \circ r)$ *is not valid.*

Proof. Right-to-left distribution. For any w, π, $(m + n)(w, \pi) = m(w, \pi) \cup n(w, \pi)$ (Definition 2). Now for any π', let $\pi = r(w, \pi')$: we have $(m + n)(w, r(w, \pi')) = m(w, r(w, \pi')) \cup n(w, r(w, \pi'))$. Thus by Definition 2, $(m + n) \circ r = m \circ r + n \circ r$.

Counterexample to left-to-right distribution. Write $p \vee q$ the disjunction of any propositions p and q. Let m be such that $m(w, \pi) = \{p \vee q : p, q \in \pi\}$. Consider w, n, r such that $n(w, \emptyset) = \{p\}$ and $r(w, \emptyset) = \{q\}$. Assuming $p \neq q$, we have $m \circ (n + r)(w, \emptyset) = \{p \vee q, p, q\} \neq (m \circ n) + (m \circ r)(w, \emptyset) = \{p, q\}$. □

Composition distributes right-to-left but not left-to-right because the methods algebra represents information flow or informational dependencies. Composing m with $n+r$ means that m can use the outputs of n and r together; this is not the same as applying m to the outputs of r and and those of n separately. So typically, $m \circ (n + r) \neq (m \circ n) + (m \circ r)$. By contrast, pooling together the m- and n-inferences and applying them to a single output is the same as applying m and n separately to that output, so $(m + n) \circ r = (m \circ r) + (m \circ r)$.

To sum up, we have an algebra $\langle \mathbf{M}, +, \circ, 0, 1 \rangle$ with two distinguished elements, the empty method (identity element for $+$) and the identity method (identity element for $\circ$). $+$ is associative, commutative and idempotent, $\circ$ is associative. $\circ$ distributes right-to-left over $+$ but not left-to-right.

8 Appendix B. Comparison with neighbourhood models

Methods models amount to building a neighbourhood model with two modalities out of the agent's basic methods and a background alethic modality. They are richer than simple neighbourhood models because they give insight on a structure of methods (built out of composition and union) which is, so to speak, the scaffolding with which the neighbourhood functions for knowledge and belief are built. That is why our models are more explanatory, as we will see.

A neighbourhood frame $\mathcal{F}$ is a pair $\langle W, N\rangle$ where W is a set of worlds and the $N \subseteq W \times \mathcal{PP}(W)$ a function from worlds to sets of propositions.[43]

Definition 22. Let $\mathcal{L}_\nabla$ be the set of formulas given by:

$\phi ::= \mathsf{p}|\top|\neg\phi|\phi \vee \psi|\nabla\phi$ where $\mathsf{P} = \{\mathsf{p}, \mathsf{q}, \ldots\}$ is a set of propositional constants.

Let $\mathcal{M} = \langle \mathcal{F}, V\rangle$ be a model where $V : \mathsf{P} \to P$ is a valuation function. We define $[\![\cdot]\!]^{\mathfrak{M}}$:

$[\![\mathsf{p}]\!]^{\mathfrak{M}} = V(\mathsf{p})$,

$[\![\neg\phi]\!]^{\mathfrak{M}} = W \backslash [\![\neg\phi]\!]^{\mathfrak{M}}$

$[\![\phi \vee \psi]\!]^{\mathfrak{M}} = [\![\phi]\!]^{\mathfrak{M}} \cup [\![\psi]\!]^{\mathfrak{M}}$

(and as usual for other logical connectives)

$[\![\nabla\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in N(w)\}$.

Truth. $\models^{\mathfrak{M}}_w \phi$ iff $w \in [\![\phi]\!]^{\mathfrak{M}}$.

Validity. $\models^{\mathfrak{M}} \phi$ iff for any world w, $\models^{\mathfrak{M}}_w \phi$.

In methods models, methods are functions from worlds to inference transitions functions, that is functions from sets of premises to sets of conclusions. For a given set of premises π and a given method m, the function $w \mapsto m(w, \pi)$ is a function from worlds to sets of conclusions. If propositions are sets of possible worlds, this is a neighbourhood function. In particular, the unconditional output function $w \mapsto m(w)$ of a method m is a neighbourhood function.[44] And so are the functions $B(m, w)$, $K(m, w)$, $B(w)$ and $K(w)$ that we build out of them. We can establish two useful equivalence results:

[43] Neighbourhood models (or "Scott-Montague models") have been independently introduced by Scott (1970) and Montague (1968, 1970) and explored in detail by Segerberg (1971). See Chellas' (1980, III) handbook for an overview of the results.

[44] Recall that $m(w)$ abbreviates $m(w, \emptyset)$.

Theorem 20. *For each method frame $\mathfrak{F}$, there is a pointwise equivalent neighbourhood frame $\mathcal{F}$ for the* B *operator in the simple language $\mathcal{L}_\nabla$, and conversely.*

Proof. Let $\mathfrak{F} =< W, M^B, R >$ be any methods frame. Define the neighbourhood frame $\mathcal{F} =< W, N >$ such that for any w, $N(w) = B(w)$ and for any model $\mathcal{M}$ in $\mathcal{F}$, $[\![\mathsf{B}\phi]\!]^{\mathcal{M}} = \{w : [\![\phi]\!]^{\mathcal{M}} \in N(w)\}$. We prove that $\mathcal{F}$ is pointwise equivalent to $\mathfrak{F}$ by induction on the complexity of ϕ. The interesting case is:

$\models^{\mathcal{M}}_w \mathsf{B}\phi$ iff $[\![\phi]\!]^{\mathcal{M}} \in N(w)$ (semantics)
iff $[\![\phi]\!]^{\mathcal{M}} \in B(w)$ (definition of N)
iff $[\![\phi]\!]^{\mathfrak{M}} \in B(w)$ (inductive hypothesis)
iff $\models^{\mathfrak{M}}_w \mathsf{B}\phi$.

Conversely let $\mathcal{F} =< W, N >$ be any neighbourhood frame for B. Define the methods frame $\mathfrak{F} =< W, M^B, R >$ such that $M^B = \{m\}$ where m is such that for any w, π: $m(w, \pi) = N(w)$ and R some reflexive accessibility relation. We first prove that $M = \{m\}$: $m + m = m$ and $m \circ m = m$ (the first holds for any method by the definition of union (Definition 2), the second holds because the definition of m entails that $m(w, m(w, \pi)) = N(w) = m(w, \pi)$ for any π), and since $M^B = \{m\}$, $M = M^{B\circ+} = \{m\}$ (Definition 6). We then prove that $B(w) = N(w)$: by the definition of m and Definition 8, $B(m, w) = m(w) = N(w)$ for any w. Since $M = \{m\}$, by Definition 8 again $B(w) = N(w)$. From this $\mathcal{F}$ and $\mathfrak{F}$ are easily shown to be pointwise equivalent. □

Theorem 21. *Call a neighbourhood frame $\mathcal{F} =< W, N >$ truthful iff for each w, $w \in \bigcap N(w)$.*[45] *For each methods frame $\mathfrak{F}$, there is a pointwise equivalent truthful neighbourhood frame for the* K *operator in the simple language $\mathcal{L}_\nabla$, and conversely.*

Proof. Let $\mathfrak{F} =< W, M^B, R >$ be any methods frame for K. Define the neighbourhood frame $\mathcal{F} =< W, N >$ such that for any w, $N(w) = K(w)$ and for any model $\mathcal{M}$ in $\mathcal{F}$, $[\![\mathsf{K}\phi]\!]^{\mathcal{M}} = \{w : [\![\phi]\!]^{\mathcal{M}} \in N(w)\}$. We prove as before that for any $\mathcal{M}$ in $\mathcal{F}$ and $\mathfrak{M}$ in $\mathfrak{F}$, $\models^{\mathcal{M}}_w \mathsf{K}\phi$ iff $\models^{\mathfrak{M}}_w \mathsf{K}\phi$ and that the frames are pointwise equivalent. Moreover, we prove that for any w, $w \in \bigcap N(w) = \bigcap K(w)$:

For any w, $w \in \bigcap K(w)$ iff $\forall m \forall p(p \in K(m, w) \rightarrow w \in p)$ (Definition 8).
For any w, m, p, if $p \in K(m, w)$ then $m \in M^I$ and $p \in m(w)$ (Definition 8),
if $m \in M^I$ then $p \in m(w) \rightarrow w \in p$ (Definition 7 and reflexivity of R)
Thus for any w, m, p, if $p \in K(m, w)$ then $w \in p$. So $w \in \bigcap K(w)$ for any w.

[45] The class of truthful neighbourhood frames is the class of neighbourhood frames which validate the schema $\nabla\phi \rightarrow \phi$. See Chellas (1980, 224).

Conversely, let $\mathcal{F} =< W, N >$ be any neighbourhood frame such that for any w, $w \in \bigcap N(w)$. Define the methods frame $\mathfrak{F} =< W, M^B, R >$ such that $M^B = \{m\}$ where m is such that $m(w, \pi) = N(w)$ for any w, π and R is identity. We prove that $M = \{m\}$ and $B(m, w) = N(w)$ as before. Moreover, we prove that for any w, $m \in M^I(w)$:

For any w: $m \in M^I(w)$ iff for any w', p', $wRw' \rightarrow (p' \in m(w') \rightarrow w' \in p')$ (Definition 7),

iff for any p', $p' \in m(w) \rightarrow w \in p'$ (R is identity),

iff $w \in \bigcap m(w)$,

iff $w \in \bigcap N(w)$ (definition of m), which is true by assumption.

Thus $m \in M^I(w)$ for any w. Since $M = \{m\}$, it follows that $K(w) = K(m, w) = N(w)$. From this we show as before that $\mathfrak{F}$ and $\mathcal{F}$ are pointwise equivalent with respect to K. □

The results mean that the B and K schemas valid in the class of methods frames are just those valid in the class of neighbourhood frames and in the class of truthful neighbourhood frames, respectively (see section 5.1.2).

Given the equivalences of Theorems 20 and 21, why prefer methods models to simpler neighbourhood ones? Essentially, because methods models allows us to *derive* a set of facts that would be treated as primitive in a simple neighbourhood semantics models. A simple example: despite the equivalences Theorems 20 and 21, the class of methods frames is *not* the class of neighbourhood frames for two modalities $\langle W, N^B, N^K \rangle$ where the second neighbourhood function is truthful. For it is easy to see that in our models, $p \in K(w) \rightarrow p \in B(w)$ for any w, p (Theorem 2), while we can construct neighbourhood models such that $p \in N^K(w) \wedge p \notin N^B(w)$ for some w. Of course we could introduce the notion of *belief-knowledge neighbourhood frames* $\langle W, N^B, N^K \rangle$ such that at any w, $N^K(w) \subseteq N^B(w)$ (knowledge entails belief) and $w \in \bigcap N^K(w)$ (knowledge entails truth). But that would amount to treating those facts as unexplained primitives. By contrast, those constraints on knowledge are derived in methods models from a definition of knowledge.

The same goes for other axioms. A much-discussed axiom for knowledge is **4**, according to which knowing is knowing that one knows: $Kp \rightarrow KKp$. A neighbourhood frame validates **4** iff: $p \in N^K(w) \rightarrow \{w' : p \in N^K(w')\} \in N^K(w)$ for any p, w. The condition is a transparent restatement of axiom **4**: if p is among the propositions known at w, then so is the proposition that holds wherever p is among the propositions known. The condition on the model does

not shed any light on whether, when or why the axiom should hold. Thus we are left to decide directly on the basis of the axiom whether we think our agents would or should satisfy it. By contrast, in methods models, the axiom is derived from the psychological model of the agent and the transitivity of background alethic modality. If on the relevant sense of possibility, what is possibly possible is possible, we show that the agent satisfies axiom **4** for knowledge if it has a "confident introspection" method m^{pc} such that, in non-formal terms: if she believes that p out of m then she believes that she knows that p on the basis of a composition of m and m^{pc} (section 5.4). This gives a better grasp on how and when an agent is able to know that she knows. For a start, it shows that it is not a trivial affair: the axiom fails for agents who do not introspect or if the space of possibilities is non-transitive.

The explanatory advantages of methods models are due to the fact that they contain more structure than neighbourhood ones. The neighbourhood functions B and K are not given as primitives, but constructed out of a set of methods. The construction gives us an insight into a structure of B and K that is *not* simply reducible to the structure of the set of propositions they map to (the structure of $\{p : p \in B(w)\}$ and $\{p : p \in K(w)\}$ at each w). The additional structure is reflected in the methods operators $\mathsf{B}\mu : \phi$ and $\mathsf{K}\mu : \phi$. For instance, we get validities such as:

$$\models^{\mathfrak{M}} \mathsf{K}\mu : (\phi \to \psi) \to (\mathsf{K}\nu : \phi \to \mathsf{Km}^{\mathsf{D}} \circ (\mu + \nu) : \psi)$$

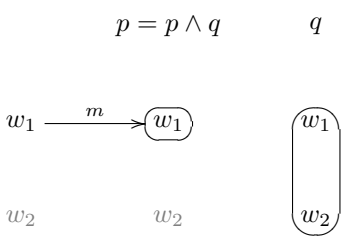
$$\models^{\mathfrak{M}} \mathsf{K}\mu : \phi \to \mathsf{Km}^{\mathsf{pi}} \circ \mu : \mathsf{B}\phi$$

which cannot be stated with the unary operators.

9 Appendix C. Counterexamples to M, N, K and 4

Example 1. Counterexample to $\mathbf{M_B}$ and $\mathbf{M_K}$. We construct a model where the agent believes and knows that $p \wedge q$, but does not believe nor know that q. Consider a frame $\mathfrak{F} = \langle W, M^B, R\rangle$ where $W = \{w_1, w_2\}$ and $M^B = \{m\}$ where m is such that $m(w_1, \pi) = \{w_1\}$ and $m(w_2, \pi) = \emptyset$ for any π, and R any reflexive accessibility relation. Let $p = \{w_1\}$ and $q = \{w_1, w_2\}$. Consider $\mathfrak{M}$ in

$\mathfrak{F}$ such that $V(\mathsf{p}) = p$ and $V(\mathsf{q}) = q$. It is easy to check that $M = \{m\}$ and that $\models^{\mathfrak{M}}_{w_1} \mathsf{B}(\mathsf{p} \wedge \mathsf{q})$ but $\not\models^{\mathfrak{M}}_{w_1} \mathsf{Bq}$, and similarly for K since however R is defined, the method m is infallible at w ($m \in M^I(w)$).

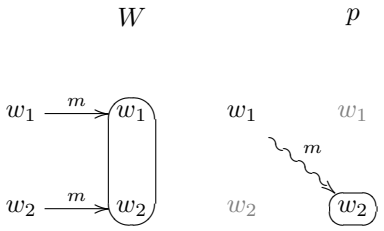


(The illustrations are explained p. 14 above.)

The methods frames $\mathfrak{F}$ constructed from a neighbourhood frame $\mathcal{F} = \langle W, N \rangle$ following the procedure in Theorem 21 are such that the agent's unique method is infallible and so they validate $\mathsf{K}\phi \leftrightarrow \mathsf{B}\phi$. (They are "excellent agent frames", in our terminology: see Definition 20). Consequently the counterexamples to the K schemas built this way, like Example 1, all involve a failure of belief, and a failure of the corresponding B schema. However they are two ways for knowledge to fail in methods models: failure of belief, but also fallibly-based belief. It will be instructive to look at two examples of the latter.

Example 2. Counterexample to $\mathbf{N}_{\mathsf{K}}$. The agent has a method that leads her to believe both the tautology and a false proposition. Though the agent believes the tautology (W), she fails to know it, because her belief is fallibly based.

Consider $\mathfrak{F} = \langle W, M^B, R \rangle$ where $W = \{w_1, w_2\}$, with $p = \{w_2\}$, and $M^B = \{m\}$ where m is s.th. $m(w_1, \pi) = \{W, p\}$ and $m(w_2, \pi) = \{W\}$ for any π; R is any reflexive accessibility relation. Consider $\mathfrak{M}$ in $\mathfrak{F}$ such that $V(\mathsf{p}) = p$. Since $p \in m(w_1)$ but $w_1 \notin p$, $m \notin M^I(w_1)$. For this it follows that $\models^{\mathfrak{M}}_{w_1} \mathsf{B}\top$ but $\not\models^{\mathfrak{M}}_{w_1} \mathsf{K}\top$ (Definitions 8 and 10).

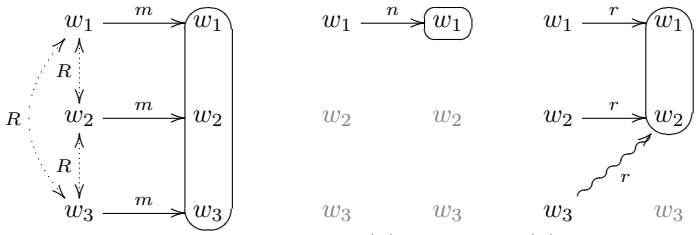


Example 3. Counterexample to $\mathbf{K}_{\mathsf{K}}$. We consider an agent that believes $p \rightarrow q$ and p out of infallible methods. The agent also believes q, but on the basis of a

method that would lead her to believe q even if it was false, so the agent fails to know that q.

Consider $\mathfrak{F} =< W, M_B, R >$ where $R = W \times W$ and $W = \{w_1, w_2, w_3\}$ with $p = \{w_1\}$ and $q = \{w_1, w_2\}$. $M^B = \{m, n, r\}$ such that $m(w, \pi) = W$ for any w, π, $n(w_1, \pi) = \{p\}$ and $n(w_2, \pi) = n(w_3, \pi) = \emptyset$ for any π, and $r(w, \pi) = \{q\}$ for any w, π.

$p \rightarrow q$ $\qquad$ p $\qquad$ q



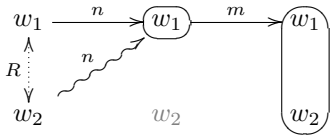
Consider $\mathfrak{M}$ in $\mathfrak{F}$ such that $V(\mathsf{p}) = p$ and $V(\mathsf{q}) = q$. We have $[\![\mathsf{p} \rightarrow \mathsf{q}]\!]^{\mathfrak{M}} = W$. Since $q \in r(w_3)$ but $w_3 \notin q$, and since $w_1 R w_3$, $r \notin M^I(w_1)$ (by Definition 7). It follows that at w_1 we have: $\models^{\mathfrak{M}}_{w_1} \mathsf{K}(\mathsf{p} \rightarrow \mathsf{q})$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Kp}$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Bq}$ and yet $\not\models^{\mathfrak{M}}_{w_1} \mathsf{Kq}$.

Example 3 models our *Watson* case. Though the agent knows two propositions that together entail q (namely $p \rightarrow q$ and p), and though she believes that q, her belief that q is not based on deduction from the two others; rather, it is has an independent and unreliable basis that would lead her to still believe that q without knowing that p and $p \rightarrow q$, and even if q was false.

Example 4. Counterexample to $\mathbf{4_K}$. The agent has a method that leads her to believe that she knows that p, even at a world where she does not.

Consider $\mathfrak{F} = \langle W, M^B, R \rangle$ where $W = \{w_1, w_2\}$, with $p = \{w_1, w_2\} = W$, $q = \{w_1\}$ and $M^B = \{m, n\}$ where m is s.th. $m(w_1, \pi) = \{p\}$ and $m(w_2, \pi) = \emptyset$ for any π, and n is s.th. $n(w, \pi) = \{q\}$ for any w, π, and $R = W \times W$.

$q = Kp$ $\qquad$ p



Consider $\mathfrak{M}$ in $\mathfrak{F}$ such that $V(\mathsf{p}) = p$. Since $p \in m(w_1)$ and $m \in M^I$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Kp}$. Since $[\![\mathsf{Kp}]\!]^{\mathfrak{M}} = q$ and $q \in n(w_1)$, $\models^{\mathfrak{M}}_{w_1} \mathsf{BKp}$. But since $w_1 R w_2$, $q \in n(w_2)$ and $q \notin w_2$, $n \notin M^I(w_1)$, and since there is no other method r such

that $q \in r(w_1)$, we have $\not\models^{\mathfrak{M}}_{w_1} \mathsf{KKp}$. (Definitions 8 and 10).

10 Appendix D. Further exploration of Reasoning methods

Axioms **M** and **C** for belief and knowledge follow from axiom **K**. But it is also possible to get them separately, by defining the two following methods:

Definition 23. *Single-Premise Deduction* is the method m^{SD} s.th $m^{SD}(w,\pi) = \{p : \exists q \in \pi(q \subseteq p)\}$ for any w, π.

Conjunctive Deduction is the method m^{CD} such that $m^{CD}(w,\pi) = \{p : \exists q, r \in \pi(p = q \cap r)\}$ for any w, π.

Theorem 22. *For any methods model $\mathfrak{M}$, if $m^{SD} \in M$, then $(\boldsymbol{M}_B) \models^{\mathfrak{M}} \mathsf{B}(\phi \wedge \psi) \rightarrow (\mathsf{B}\phi \wedge \mathsf{B}\psi)$ and $(\boldsymbol{M}_K) \models^{\mathfrak{M}} \mathsf{K}(\phi \wedge \psi) \rightarrow (\mathsf{K}\phi \wedge \mathsf{K}\psi)$ for any ϕ, ψ.*

For any methods model $\mathfrak{M}$, if $m^{CD} \in M$, then $(\boldsymbol{C}_B) \models^{\mathfrak{M}} (\mathsf{B}\phi \wedge \mathsf{B}\psi) \rightarrow \mathsf{B}(\phi \wedge \psi)$ and $(\boldsymbol{C}_K) \models^{\mathfrak{M}} (\mathsf{K}\phi \wedge \mathsf{K}\psi) \rightarrow \mathsf{K}(\phi \wedge \psi)$.

Proof. The proofs are analogous the proof of $\mathbf{K}_B$ and $\mathbf{K}_K$ (Theorem 7).

For $(\mathbf{M}_B)$ and $(\mathbf{M}_K)$, assume that $\mathfrak{M}, w$ are s.th. $\models^{\mathfrak{M}}_w \mathsf{B}(\phi \wedge \psi)$ and that $m^{SD} \in M$. Then there are p, q such that $[\![\phi]\!]^{\mathfrak{M}} = p$, $[\![\psi]\!]^{\mathfrak{M}} = q$, and $p \cap q \in m(w)$ for some $m \in M$. We show that $p, q \in (m^{SD} \circ m)(w)$, that $m^{SD} \circ m$ is infallible if m is, and that $m^{SD} \circ m \in M$. From this $(\mathbf{M}_B)$ and $(\mathbf{M}_K)$ follow.

For $(\mathbf{C}_B)$ and $(\mathbf{C}_K)$, assume that $\mathfrak{M}, w$ are s.th. $\models^{\mathfrak{M}}_w \mathsf{B}\phi \wedge \mathsf{B}\psi$ and that $m^{CD} \in M$. Then there are p, q such that $[\![\phi]\!]^{\mathfrak{M}} = p$, $[\![\phi]\!]^{\mathfrak{M}} = q$, and $p \in m(w)$ and $q \in n(w)$ for some m, n in M. We show that $p \cap q \in (m^{CD} \circ (m+n))(w)$, that $m^{CD} \circ (m+n)$ is infallible if m and n are, and that $m^{CD} \circ m \in M$. From this $(\mathbf{C}_B)$ and $(\mathbf{C}_K)$ follow. □

The relation between Multi-Premise Deduction, Single-Premise Deduction and Conjunctive Deduction is straightforward:

Corollary 4. $m^D = m^{SD} \circ m^{CD}$.

Proof. Evident from Definitions 12 and 23. □

In neighbourhood models, axioms **M**, **C** and **K** have been correlated to corresponding properties of the topology of sets of sets (see Chellas, 1980, 215–216). A set of sets $S \subseteq \mathcal{PP}(W)$ is *supplemented* or closed under supersets iff

$\forall X, Y \subseteq W((X \in S \wedge X \subseteq Y) \rightarrow Y \in S)$, is *closed under finite intersections* iff $\forall X, Y \in S(X \cap Y \in S)$, *contains its core* iff $\bigcap S \in S$, and is *augmented* iff it is supplemented and contains its core. We say that a neighbourhood function $W \rightarrow \mathcal{PP}(W)$ is supplemented, closed under finite intersections, and augmented iff it maps to supplemented, closed under finite intersections, and augmented sets, respectively. Supplemented neighbourhood functions satisfy **M**, neighbourhood functions that are closed under finite intersections satisfy **C**, and augmented ones satisfy **K**. It can be shown that the methods m^{SD}, m^{CD} and m^{D} ensure that the B and K neighbourhood functions are respectively supplemented, closed under finite intersections, and augmented. We do not present the proofs here. The reader will easily construe them by considering, for a given method m, the series of composed methods m^{Dm_k}, $k \in \mathbb{N}$ such that $m^{Dm_0} = m$, and $m^{Dm_k} = m^{D} \circ m^{Dm_{k-1}}$ for any $k \geq 1$, and analogous series for m^{SD} and m^{CD}.

Satisfying the knowledge axioms $\mathbf{M}_K$, $\mathbf{C}_K$, $\mathbf{K}_K$ does not guarantee satisfaction of the corresponding belief axioms, and conversely. It is in principle possible that an agent believes the logical consequences of what she believes on the basis of infallible methods but does not believe all the logical consequences of what she believes on the basis of fallible methods.

Finally, note that we have only stated *sufficient* conditions for a methods frame to satisfy **M**, **C**, **K** and **N**, not necessary ones. There are frames that validate those schemas for belief and/or knowledge without Deduction and Pure Reason. Consider two examples:

- $M^B = \{m^R\}$. The agent believes and knows the tautology, and only the tautology, at any world. Trivially, the agent validates $\mathbf{K}_B$ and $\mathbf{K}_K$, yet she does not have m^D.

- Suppose that an agent is such that whenever, for some w, m, n, $p \in B(m, w)$ and $(W \backslash p) \cup q \in B(n, w)$, there is some third method r such that $q \in B(r, w)$, yet the third method is not the result of composing m and n with m^D. For instance, one may assume that r *also* outputs some true propositions (say, $\{w\}$ at any w) that do not follow from the outputs of m and n. Such an agent can satisfy $\mathbf{K}_B$ and/or $\mathbf{K}_K$ without having m^D.

Therefore the methods m^R and m^D fail to identify the class of methods frames that validate the schema of normal modal logics (**KN**). By contrast, in neighbourhood semantics, these class of frames can be identified as the ones in which

the neighbourhood function is augmented. Is that a defect of our models? Quite the contrary. An agent may satisfy the **KN** schemas "accidentally", so to speak. Imagine an agent that forms beliefs by listening to various people's testimonies. Suppose that whenever the agent has heard p and $p \rightarrow q$ from some persons, there happens to be, by sheer coincidence, a person that tells her that q. The agent thereby satisfies $\mathbf{K}_B$, $\mathsf{B}(\mathsf{p} \rightarrow \mathsf{q}) \rightarrow (\mathsf{Bp} \rightarrow \mathsf{Bq})$. Yet that is intuitively accidental, because her belief that q is unconnected to her believing p and $p \rightarrow q$. By contrast, it is not all accidental that an agent satisfies $\mathbf{K}_B$ if the agent has the Deduction method, because the method ensures that she has a belief that q based on her having beliefs that p and that $p \rightarrow q$. The upshot is that, far from being a deficiency of methods frames, the fact that there is no natural class of methods frames that validates **KN** rather shows that the epistemic and doxastic **KN** are superficial rather than deep generalities about knowledge and belief. For instance, the deep generality behind **K** is that a (certain idealised type of) agent *deduces* all the logical consequences of what she knows, and that generality can only be stated in the more complex language that allows reference to methods (Definition 10), as we pointed out (see Theorem 8).

11 Appendix E. Belief, knowledge and information

Given methods models, we can define the information provided by a method as the set of possibilities compatible with its outputs. Correlatively, we define an agent's doxastic and epistemic information as the set of possibilities compatible with what an agent believes and with what she knows.

Definition 24. For any method m, $I(m,w) = \bigcap\{p : p \in m(w)\}$ is the doxastic information given by method m, and $E(m,w) = \bigcap\{p : p \in m(w) \wedge m \in M^I(w)\}$ is the method's epistemic information. (If the method is infallible, epistemic and doxastic information coincide; if the method is fallible, its epistemic information is nihil.)

$I(w) = \bigcap B(w)$ is the agent's doxastic information.

$E(w) = \bigcap K(w)$ is the agent's epistemic information.

The doxastic and epistemic information of a *set* of methods X is the conjunction of the information of its members: $I(X,w) = \bigcap_{m \in X} I(m,w)$ and $E(X,w) = \bigcap_{m \in X} E(m,w)$.

As understood here, information is an objective notion of content. The doxastic information provided by a method is how the world has to be in order to be as the method presents it: that is, which states of the world are compatible with the output of a method. Epistemic information is the information provided by knowledge-producing methods: how the world has to be in order to be as a method teaches that it is.

The formal representation of such notions is familiar from Hintikka (1962) models. The states of worlds that are compatible with what I believe are just those worlds where every proposition I believe to be true holds. So at a given world w, the set of worlds is the intersection of the set of propositions in B_w. And similarly for K_w or for the unconditional outputs of a given (infallible) method m. Our notions of information thus correspond to standard Kripke models:

Remark 6. For a frame $\mathfrak{F}$ and any method m, let $R^{Im}, R^{Em}, R^{I}, R^{E} \subseteq W \times W$ be such that, for any w, w':

$wR^{Im}w'$ iff $w' \in I(m, w)$,
$wR^{Em}w'$ iff $w' \in E(m, w)$
$wR^{I}w'$ iff $w' \in I(w)$
$wR^{E}w'$ iff $w' \in E(w)$
Each of $\langle W, R^{Im}\rangle$, $\langle W, R^{Em}\rangle$, $\langle W, R^{I}\rangle$ and $\langle W, R^{E}\rangle$ is a Kripke frame.

The relations between $I(m, w)$, $E(m, w)$, $I(w)$ and $E(w)$ are straightforward. If m is infallible at w, $E(m, w) = I(m, w)$, otherwise $E(m, w) = \emptyset$. It is easy to check that $I(w) = \bigcap_{m \in M} I(m, w)$ at any w (where M is the agent's method set), and that $E(w) = \bigcap_{m \in M \cap M^I(w)} I(m, w)$ at any w, where $M \cap M^I(w)$ is the set of agent's methods that are infallible at w.

Intuitively, a method is informative iff it can reduce the set of possibilities the agent considers or should consider. Formally, we can formulate the intuition in two ways, depending on whether the method is inferential or non-inferential:

Definition 25. A purely non-inferential method m is *informative* iff for some w, n, $I(n, w) \neq I(m + n, w)$.[46]

An inferential or mixed method m is informative iff for some w, n, $I(n, w) \neq I(m \circ n, w)$.

Non-informative methods are guaranteed to preserve truth, in the following sense:

[46] A method m is purely non-inferential iff $m \circ n = m$ for any n (section 2.2).

Corollary 5. *Say that the information of m is truthful at w iff $w \in I(m, w)$. If an inferential method m is non-informative, then for any n, $I(m \circ n, w)$ is truthful if $I(n, w)$ is. If a non-inferential method m is non-informative, then for any n, $I(m + n, w)$ is truthful if $I(n, w)$ is.*

Proof. Obvious from Definition 25. □

This characterises the sense in which non-informative methods are risk-free. Applying an inferential non-informative method to another one cannot lead to falsity unless the original method was erroneous; adding a non-inferential non-informative method to any other cannot lead to a false set of beliefs unless the original one did.

Theorem 23. *Deduction and Pure-Reason are not informative. For any m, w, $I(m, w) = I(m + m^R, w)$ and $I(m, w) = I(m^D \circ m, w)$.*

Proof. By Definition 11, at every w, $m^R(w) = \{W\}$. Since for any p, $p \in m(w) \rightarrow p \subseteq W$, $I(m, w) = \bigcap m(w) \subseteq W$. So $I(m + m^R, w) = \bigcap(m(w) \cup \{W\}) = \bigcap m(w) = I(m, w)$.

For Deduction, let $o = \bigcap m(w)$ for some w, m. For any $q, r \in m(w)$, $q \cap r \supseteq o$. So for any $p \supseteq q \cap r$, $p \supseteq o$, so that $o \cap p = o$. From this and Definition 12 it follows that for any $p \in m^D \circ m(w)$, $p \supseteq o$. So $\bigcap(m^D \circ m)(w) \supseteq \bigcap m(w) = o$. Moreover, since for every p, $p \subseteq p \cap p$, we have $m(w) \subseteq m^D \circ m(w)$, so that $\bigcap(m^D \circ m)(w) \subseteq \bigcap m(w)$. □

Given any set of beliefs, Pure Reason adds belief in the proposition $\{W\}$, so it cannot narrow down the set of worlds compatible with the beliefs. And given any premises, Deduction adds supersets of intersections of them, so it cannot narrow down the set of worlds compatible with the intersection of all of them.

The result formalises the intuition that Reason and Deduction do not provide information that was not at least implicit in the premises, and the intuition that they are risk free methods.

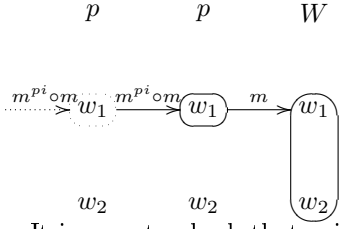
Conjecture 1. *Say that a method m is included in a method n iff for all w, π, $m(w, \pi) \subseteq n(w, \pi)$. Let Total Deduction be such that $m^{TD} \circ n = m^D \circ m^D \circ ... \circ m^D \circ n$.*

Reason and Total Deduction are the most inclusive non-informative methods. Any purely non-inferential method that is not included in Reason is informative; any purely inferential method that is not included in Total Deduction is informative.

By contrast, Introspection methods are potentially informative. Most interestingly, they are potentially informative with respect to the method or sets of methods they introspect:

Theorem 24. *Introspection and Confident Introspection are informative. For some methods models, there are* $w, m, X, pi(m), pc(m), ni(X), nc(X)$ *(where* $X \subseteq \mathbf{M}$ *is a set of methods) such that* $I(m,w) \neq I(\{m \cup pi(m)\}, w)$*,* $I(m,w) \neq I(\{m \cup pc(m)\}, w)$*,* $I(X,w) \neq I(X \cup \{ni(X)\}, w)$*, and* $I(X,w) \neq I(X \cup \{nc(X)\}, w)$*.*

Proof. Model for $I(\{m \cup pi(m)\}, w) \neq I(m,w)$. Let $W = \{w_1, w_2\}$ and $M^B = \{m, n\}$ where m and n are purely non-inferential methods such that: $m(w_1) = \{W\}$, $n(w_1) = \{w_1\}$ and $m(w_2) = n(w_2) = \emptyset$. Since m and n are purely non-inferential, we have $M = \{m, n, m+n\}$.



It is easy to check that n is the Positive Introspection of m for this model: $n = pi(m)$ (Definition 15). At w_1, m outputs W and only W. Correspondingly, n outputs $b_W = \{w : \exists m' \in M(W \in m'(w))\} = \{w_1\}$. At w_2, m outputs nothing and the output of n is correspondingly empty. At w_1 we have $I(m,w) = W$ but $I(m,w) \cap I(pi(m),w) = W \cap w_1 = w_1$.

Similar models can be built for $I(m,w) \neq I(\{m \cup pc(m)\}, w)$, $I(X,w) \neq I(X \cup \{ni(X)\}, w)$, and $I(X,w) \neq I(X \cup \{nc(X)\}, w)$. □

The results are fairly intuitive. Typically, p and Bp do not hold at the same worlds. For that reason, an Introspection method that "adds" a belief that Bp wherever the agent believes p typically narrows down the sets of worlds compatible with the agent's beliefs. Similarly, p and Kp do not typically hold at the same worlds. That is why Introspection and Confident Introspection are informative methods. Because they narrow down the set of worlds compatible with the agent's beliefs, they are not risk-free methods: it may be true that p and false that Bp, or true that p and false that Kp. They are inductive methods, as we defined them section 3.4. Correlatively, the axioms of epistemic logic that rely on them (**4** and **5**) are not a matter of pure rationality or inner coherence; they require reliable information-gathering methods.

References

David M. Armstrong. *Belief, Truth and Knowledge*. Cambridge University Press, 1973.

Sergei Artemov. The logic of justification. *The Review of Symbolic Logic*, 1: 477–513, 2008.

Sergei Artemov. Logic of proofs. *Annals of Pure and Applied Logic*, 67:29–59, 1994.

Sergei Artemov and Elena Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15:1059–1073, 2005.

Brian F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

Roderick M. Chisholm. *Theory of Knowledge*. Prentice Hall, Englewood Cliffs, NJ, 1966. ISBN 0139141502.

Michael Clark. Knowledge and grounds: A comment on mr. gettier's paper. *Analysis*, 24(2):46–48, 1963. URL http://www.jstor.org/stable/3327068.

Keith DeRose. Solving the skeptical problem. *The Philosophical Review*, 104 (1):1–52, 1995. URL http://www.jstor.org/stable/2186011.

Fred Dretske. Conclusive reasons. *Australasian Journal of Philosophy*, 49:1–22, 1971.

Ronald Fagin and Joseph Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34(1):39–76, 1988. doi: 10.1016/0004-3702(87)90003-8.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

Melvin Fitting. Justification logics, logics of knowledge, and conservativity. *Annals of Mathematics and Artificial Intelligence*, 53:153–167, 2008.

Melvin Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132:1–25, 2005.

Gottlob Frege. On sense and reference. In Peter Geach and Max Black, editors, *Translations from the Philosophical Writings of Gottlob Frege*. Blackwell, 1892/1980.

Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963. URL http://www.jstor.org/stable/3326922.

Anthony Gillies. Counterfactual scorekeeping. *Linguistics and Philosophy*, 30: 329–360, 2007. doi: 10.1007/s10988-007-9018-6.

Alvin I. Goldman. Discrimination and perceptual knowledge. *Journal of Philosophy*, 73(20):771–791, 1976. doi: 10.2307/2025679. URL http://www.jstor.org/stable/2025679.

John Hawthorne. *Knowledge and Lotteries*. Oxford University Press, 2004. ISBN 0199287139.

Vincent Hendricks. *Mainstream and Formal Epistemology*. Cambridge University Press, 2006.

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

Jaakko Hintikka. *Socratic Epistemology*. Cambridge University Press, 2007.

David Kaplan. Demonstratives. In Joseph Almog, John Perry, and Howard Wettstein, editors, *Themes from Kaplan*, pages 481–563. Oxford University Press, 1989.

Kevin Kelly. *The Logic of Reliable Enquiry*. Oxford University Press, 1996.

Saul Kripke. *Naming and Necessity*. Harvard University Press, 1980.

Saul Kripke. A puzzle about belief. In Avishai Margalit, editor, *Meaning and Use*. Reidel, 1979.

David Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74:549–567, 1996.

David K. Lewis. *Counterfactuals*. Harvard University Press, 1973.

William G. Lycan. On the gettier problem problem. In Stephen Hetherington, editor, *Epistemology Futures*, pages 148–169. Oxford University Press, 2006.

Richard Montague. Pragmatics. In R. Klibansky, editor, *Contemporary Philosophy: a Survey*, pages 102–122. La Nuova Italia Editrice, 1968.

Richard Montague. Universal grammar. *Theoria*, 36:373–398, 1970.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, Mass., 1981.

John Perry. Frege on demonstratives. *Philosophical Review*, 86(4):474–497, 1979.

Willard van Orman Quine. Reference and modality. In *From a Logical Point of View*, pages 139–159. Harvard University Press, 2nd edition, 1953/1961.

Dana Scott. Advice in modal logic. In K. Lambert, editor, *Philosophical Problems in Logic*, pages 143–173. Reidel, 1970.

Krister Segerberg. *An Essay in Classical Modal Logic*, volume 13 of *Filosofiska Studier*. University of Uppsala, 1971.

Ernest Sosa. Postscript to "proper functionalism and virtue epistemology". In John L. Kvanvig, editor, *Warrant in Contemporary Epistemology*. Rowman & Littlefield, Lanham, Md, 1996.

Robert Stalnaker. A theory of conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory*. Blackwell, 1968.

Jason Stanley. *Knowledge and Practical Interests*. Oxford University Press, 2005.

Scott Sturgeon. The gettier problem. *Analysis*, 53(3):156–164, 1993. URL http://www.jstor.org/stable/3328464.

Peter Unger. An analysis of factual knowledge. *Journal of Philosophy*, 65(6): 157–170, 1968. URL http://www.jstor.org/stable/2024203.

Johan van Benthem. Epistemic logic and epistemology. *Philosophical Studies*, 128:49–76, 2006.

Kai von Fintel. Counterfactuals in a dynamic context. In M. Kenstowicz, editor, *Ken Hale: A life in language*. MIT Press, 2nd edition, 2001.

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