Methods Models for Belief and Knowledge*

20th February 2014

WORK 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 and Lottery cases, inductive knowledge, fallible justification, and 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 requires 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. We additionally show how to use the models to define notions of common belief and knowledge, information and informativeness, and implicit knowledge and belief.

1 Introduction

Our 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

*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, Johannes Stern, Anne Meylan, Davide Fassio, Arturs Logins and Martin Smith for their help and encouragement, and to audiences in Geneva (Palmyr workshop), Lausanne (DLG09), Nancy (EpiConfFor workshop), Konstanz (FEW 2010), Copenhagen (Formal Epistemology Workshop) and two anonymous referees for useful comments.

in a number of applications. We introduce formal models of knowledge stemming from the idea that knowledge is a matter of bases of belief, or as we 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. Most importantly, they provide a philosophically satisfying representation of knowledge. We show that by formalising several important epistemological notions such as Gettier cases, Lottery cases, inductive knowledge, fallible justification and deductive closure without logical omniscience.

Let us first illustrate the puzzle:

1. *Tea leaves.* 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 appropriate 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 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 the 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

[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.

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

Granted, one can amend or extend epistemic logic to deal with some basis-related aspects of knowledge. Much has already be done in that respect.[4] Yet we take a different approach here. Instead of adapting standard models to specific philosophical purposes, we start from a philosophical characterisation of knowledge and build models suited to it.

Our guiding idea is that knowledge is *infallibly based belief*:

[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 one is in. The move has unpalatable consequences: see 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."

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

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

The account relies on two primitives: *methods* and some notion of *possibility* expressed by "could". *Methods* are whatever bases of belief or ways or forming or sustaining beliefs that are relevant to knowledge attribution.[5] A method "yields" a belief if it forms or sustains it; one believes that p "on the basis" of a method if that method yields a belief that p. We do not assume that each belief is based on a single method: several methods may yield the same belief. To give a few illustrations: 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.

The broad characterisation of the notion of method leaves unsettled a number of questions one may raise. They need not be conscious procedures one follows methodically. They may or may not involve unconscious computational processes. They may or may not be individuated "externally" or "broadly", that is, they may or may not involve relations between an agent and her environment and aspects of the latter. We need not adjudicate here. Our models investigate what we can say about knowledge while keeping a fairly abstract perspective on what methods are.[6]

The relevant notion of *possibility* is alethic, not doxastic or epistemic. For a method to be infallible, it is not required that one believes or knows that it is. It is sufficient that error is in fact impossible. Yet within the domain of alethic possibilities, 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). Again, we need not make a choice here; our models abstract away from the details of the relevant notion of possibility. Two points should be mentioned, though. First, the actual world is possible, in the relevant sense. Thus for a method to be infallible, it

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

[6] My own view is that there is no workable *pretheoretical* notion of methods. Rather, methods are functionally individuated by the role they play in the method-infallibilist account of knowledge. We therefore have to rely on our pretheoretical judgements about knowledge in order to find out what methods are. This does not mean that methods are ontologically less basic than knowledge; it merely means that they are theoretical entities. Thus the models can be equally seen as contributing to the understanding of methods as to that of knowledge. However, nothing here hinges on that particular view of methods.

should not actually yield a false belief. Second, possibility may be metaphysically contingent. In some worlds, pigs can fly, in others, they cannot; similarly, some methods may be infallible at some worlds and fallible at others.[7]

There are several reasons to adopt the method-infallibilist account, which we can only mention here. First, it secures factivity (section 5.1.2) and the idea that deduction preserves knowledge (see Lasonen-Aarnio, 2008 on how fallibility threatens deductive closure).[8] 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 used to represent fallibilist notions of knowledge (section3.4).

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

Assuming certain forms of essentialism, there is no possible world where *that* orange, the one Oscar is looking at, is not a fruit (Kripke, 1980). And there is no possible world 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 is a fruit too, and Primo would falsely believe on the same basis that 49 is prime. Our notion of method directly implements the idea that we evaluate a belief by looking at whether it is, so to speak, in "good" or "bad" company.

[7] My preferred view is that the relevant notion of possibility is the one that underlies the so-called "circumstantial" uses of modals (Kratzer, 1981). If such modals are context-sensitive, this would be a family of notions among which context pick up a relevant one (see section 2.3).

[8] Note that deductive closure is not equivalent to logical omniscience (section 5.2).

[9] A variant of Ginet-Goldman barn facades case (Goldman, 1976, 772–3).

Our method-infallibilist account makes one questionable assumption. We say that whether one knows in a given case depends on what goes on for beliefs with the *same* basis. Hence we have an equivalence class of beliefs based on the same basis such that either all or none constitute knowledge. One may find it more plausible to think that knowledge in a case requires avoidance of error in cases involving *similar enough* bases. For instance, whether I know that the object in front of me is an apple partly depends on whether I would misidentify it if it was slightly further away. But whether I would *know* that it is an apple if it was slightly further away depends on whether I would go wrong if it was even further away, and so on. Assuming that knowledge only depends on beliefs with the same basis would lead us to consider that we have a unique basis here, however far away the object would be. But that is implausible. However we will ignore the issue here, and leave for further work the exploration of how our models may be recast in the less demanding setting of a similarity relation between bases.[10]

Section 2 gives the most general characterisation of methods and infallibility, without assuming a specific notion of proposition. We also define the operations of union and application of methods. A corresponding algebra is given in appendix A. Section 3 applies the method-infallibilist approach to formalise Gettier cases, Lottery cases, inductive knowledge and fallible justification. Section 4 introduces methods models formally. 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. 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. Appendices D and E discusses further the Reasoning and Introspection methods we introduce to derive our results. Multi-agents seetings are also briefly discussed in Appendix E, and notions of common belief and common knowledge are defined in methods terms. Appendix F introduces notions of information and informativeness. These provide a way to relate methods models to standard Hintikka models.

[10] Thanks to John Hawthorne here.

2 Methods

2.1 The space of possible methods

A method is something that yields beliefs. But it need not always yield the same beliefs; in fact, it typically yields 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 *non-inferential* method, and it is naturally modelled as a function from worlds to sets of propositions. Perception is a paradigmatic non-inferential method. Second, a method may yield beliefs as a function of other beliefs. For instance, *modus ponens* would lead one to believe that q from the premises that p and that $p \to q$. 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). Deduction is a paradigmatic inferential method.

Generalising both ideas, a 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 are variable functions from worlds and sets of premises to sets of conclusions.

We thus define the space of all possible methods as the space of all functions from worlds and sets of premises to sets of conclusions. When a method maps to certain conclusions from a world and certain premises, we say that the method *reaches* those conclusions from those premises at that world. When the proposition p is among the conclusions a method m reaches from a set of premises π at a world w, we say that m *outputs* p *from* π at w. When p is among the conclusions which a method m reaches from the empty set of premises at w, we say that m *unconditionally outputs* p at w.

The idea that a method is a function from *worlds* is a gross simplification. We may want to say that a same method can yield different beliefs if used at different positions in space or time. Thus methods should rather be functions from *centred worlds* and premises to conclusions, where a centred world consists of a (classical) world and a *perspective* on it: a subject and a time, at least (Quine, 1969, 154, Lewis, 1979/1983, 147). We ignore the complication here. Up to a point, it may be accommodated by interpreting the "worlds" in our

models as referring to centred worlds.[11]

We take beliefs to be the unconditional outputs of the methods an agent has. Methods could be also used to cash out a notion of conditional belief, but we do not explore it here.

2.2 Operations on methods

The *union of a method m and a method n* is a method that outputs (relative to a world and a set of premises) the union of the outputs of m and n (relative to that world and that set of premises). Method union is the idea that an agent is able to pool together the outputs of different methods. Limits on method union thus reflect 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.

The *application of a method m to a method n* is the method that outputs (relative to a world and a set of premises) the conclusions that m reaches from the conclusions that n reaches (relative to that world and that set of premises). Method application is the idea that an agent is able to apply one method to the output of another. If n outputs the conclusion set $\{p\}$ from no premises, and m outputs q from the set $\{p\}$, then the application of m to n reaches q from no premises. Limits on application thus correspond to *computationally bounded* agents, such as agents who are only able to go through proofs with a limited number of steps.[12]

Method union and application are understood as synchronic operations here: they are not meant to represent dynamic processes of belief formation or revision, but rather epistemic dependencies among one's beliefs.

The two operations give an algebra over the space of methods. Its main properties are detailed in appendix A.

2.3 Infallibility

A method is infallible if and only if it is impossible for it to reach false conclusions from true premises. The fact that *modus ponens* sometimes reaches false conclusions from false premises does not make it a fallible inference rule. The *empty* set of premises is trivially such that all the premises in it are true, so

[11] Thanks to John Hawthorne here.

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

infallibility requires in particular that it is impossible for the method to have false unconditional outputs.

We model the relevant notion of possibility standardly by introducing a reflexive accessibility relation over worlds. The introduction of a relation allows for metaphyiscally contigent notions of possibility: some things may be possible at a world but impossible at another. The reflexivity ensures that the actual is possible in the relevant sense of possibility. Note that the accessibility relations represent a background notion of alethic possibility, not epistemic accessibility as in standard Hintikka models.

Thus a method is infallible at a world if and only if at any world accessible from it, for any sets of premises all of which are true at that world, the conclusions the method reaches from these premises at that world are also true at that world.

Infallibility is preserved under union and application. The union of infallible methods is itself infallible, and the application of an infallible method to an infallible one is itself infallible.[13]

Various constraints on the relation yield different notions of infallibility and correspondingly implement various modal accounts of knowledge:

1. Each world has only access to itself: 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.[14]

2. Each world has access to all worlds: a notion of knowledge that requires the metaphysical impossibility of error, and correspondingly precludes inductive knowledge. The view could be ascribed to Descartes.

3. Each world has access to a range of "close" worlds only: 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 5.4).

4. Each conversational context associates "knows" with a specific accessibility relation: a modal contextualist account along the lines of DeRose (1995) and Lewis (1996).

[13] On a weaker notion of infallibility only the unconditional outputs of methods (*i.e.* beliefs) would be taken into account. On this weaker notion application does not preserve infallibility. Thanks to Timothy Williamson for suggesting the stronger notion used here.

[14] 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.

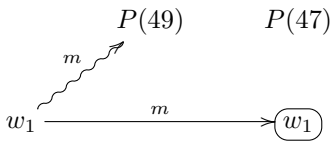
Most of our results are independent of the choice. They only require that the relation be reflexive. The exception is knowledge of one's knowledge, which we derive only on the assumption that what is possibly possible is possible — that is, by requiring a transitive accessibility relation.[15]

3 Philosophical Applications

In this section we sketch how to represent some important epistemological ideas in terms of methods models.

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 that he considered both questions at the actual world.) Let $P(47)$ and $P(49)$ be the relevant propositions. We may represent the case as follows:



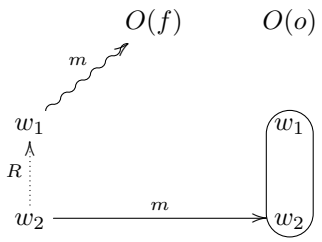
[15]Some modal accounts of knowledge cannot be represented without substantial modifications of our apparatus. On Nozick's (1981, chap.3) view, one knows that p only if: *if p had been false, one would not have believed p.* On the Lewis-Stalnaker semantics of counterfactuals (Stalnaker, 1968; Lewis, 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 possible worlds 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 may relativise accessibility to propositions. Infallibility is correspondingly proposition-relative, and we say that one knows p iff one believes p on the basis of a method that is infallible with respect to p. This invalidates our proof (section 5.2) that Deduction preserves knowledge: while the infallibility of a method with respect to p ensures the infallibility of the application of Deduction to that method *with respect to* p, it does not ensures the infallibility of the application of Deduction to that method *with respect to the conclusions reached* by this application. 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 conditional is true iff $p \rightarrow q$ holds throughout a set of close worlds fixed by context. The set does not depend on the particular p evaluated. On that semantics Nozick's condition can be modelled in our system by a context-relative accessibility relation, 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 proposition-dependent as in Nozick's semantics. So Infallibility must be relativised to propositions again, and our proofs do not go through. If stakes are only relative to a subject's situation, we can model stake-sensitivity with an accessibility relation relativised to subject and time only. On the latter model stake-relative versions of our results can be recovered.

(*Illustrations for methods models.* To avoid cluttering, I lay out the set of possible worlds in a column that is repeated horizontally as many times as needed. Propositions are represented by circled sets of worlds; these are just the worlds in which the proposition is true. Typically we use one column per proposition. Only unconditional outputs of methods — i.e. beliefs — are represented: an arrow named m between w and p indicates that p is an unconditional output of m at w. 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 false in every world, as $P(49)$ is here, the arrow points directly to its name instead of a circle of worlds. When needed, accessibility relations between worlds are represented by dotted arrows labelled 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 at w_1, and no belief based on m at w_1 can be knowledge.

Similar models can be given for the fake-barn style of case (section 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 $O(f)$ for the relevant propositions:



The only belief m yields at w_2 is true. But there is a world accessible at w_2 in which m yields a false belief, namely w_1. So the method is fallible at w_2, and that is why the subject fails to know even in world w_2.

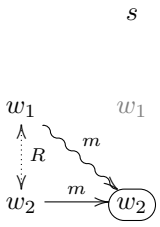
The models are straightforward but not trivial. In standard epistemic logic, Kp holds if and only if p is true at all (epistemically) accessible worlds. Unless impossible worlds are introduced, it is true at every world that 47 is prime. So however accessibility is fixed, we get the result that it is 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 w iff t is evidence for p at w and p holds at all accessible worlds.[16] Here we get the result that it is known 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

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

calculations as justifications, we get the wrong results; if we do not, the models do not explain why they do not count as justifications.[17]

3.2 Standard Gettier cases

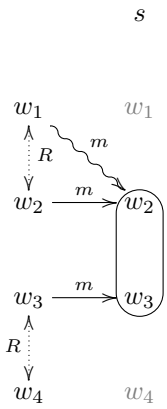
Consider (a slight variant of) Chisholm's (1966, 23n) sheep case: a man comes to believe that there is a sheep in a field by seeing a sheep-looking rock in the distance. As it happens, there is one, but it is hidden behind the rock. The case is straightforwardly modelled if we assume that the subject 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:



The actual world is w_2. At that world, the man's method outputs the belief that there is a sheep in the field, which is true. However, there is an accessible world w_2 in which the same method would yield a false belief: that is the world in which the sheep-looking rock is still present but there is no sheep behind it. So the method is fallible at the actual world and the man fails to know.

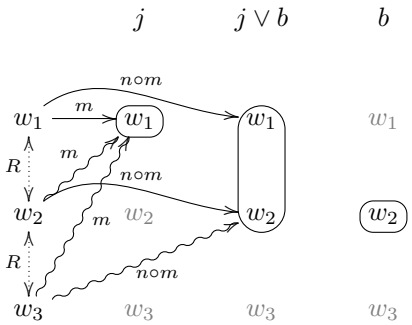
Now one may add that what is distinctive about Gettier-style cases is that the agent's failure to know is due to something abnormal in the circumstances. In normal environments the man's method would have been adequate for him to come to know that a sheep in the field. That is why the agent is said to be justified even though they do not know. On the method-infallibilist account, this implies that the fallibility of m is only metaphysically contingent. In normal worlds, the method would be infallible. Correspondingly, we may supplement our model as follows:

[17] 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.



The modal space is now divided into two cells: $\{w_1, w_2\}$ and $\{w_3, w_4\}$. The former represent the abnormal circumstances in which a sheep-looking rock is present: in those, the man's method is fallible, and he fails to know even when he is right. The latter, however, represent the normal circumstances in which no misleading rock is present. In those circumstances, the man's method yields a belief that there is a sheep only if there is in fact one. Thus the method is infallible in those worlds and yields knowledge.[18]

The original Gettier (1963) cases are essentially modelled in the same way, except that we need to introduce a inferential step. Smith has good evidence 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 leads him to infer $j \vee b$ from j. Write $n \circ m$ for the application of n to m. Assuming that the subject could have formed the same beliefs while b was false:



[18] In the model, the method is also "omniscient" at $\{w_3, w_4\}$ in the sense that whenever there is a sheep, it yields the belief that there is one. This is not required for knowledge on our account. One could add a world w_5 to the cell $\{w_3, w_4\}$ in which the sheep is present but the man does not form the belief that there is one. (The sheep may be hidden or far away, for instance.) The method would still be infallible over $\{w_3, w_4, w_5\}$ and would thus give knowledge wherever it yields belief.

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 b 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. Only abnormal circumstances are represented, but one can easily supplement the model with worlds representing normal ones in which $n \circ m$ would be infallible.

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.

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 3.3). By contrast, our models can straightforwardly implement the idea that one's justification in a Gettier case is of a kind that would provide knowledge in normal circumstances.[20]

3.3 Inductive knowledge

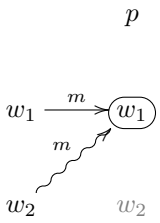
Things being as they are, my mother often comes to *know* that the neighbours are home by seeing that their light is on. Absolutely speaking, it is of course physically possible that the neighbour's light is on and they are not there. But given their habits and the general circumstances, they could not 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 on. One the method infallibilist view, that is

[19] See Lycan (2006, section 6.1) for a recent defence of the no-false-assumption view.

[20] To be clear: the method infallibilist account alone does not tell us why the error-world associated to a Gettier case is accessible from the Gettier situation but not from the normal worlds. That depends on the set-up of the accessibility relation, which is a primitive of our models. In that sense, the account does not "solve" the Gettier problem. However, it tells us how the modal space must be like provided that the case is a Gettier-style case. Namely, and error world must be accessible, and there must be some worlds (the "normal" ones) in which the agent's method would be infallible.

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.[21]

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 p is known.

Good inductive methods are thus infallible, but only locally. They are "fallible" only in the sense that their infallibility is metaphysically contingent: there are some worlds where there are not infallible. By contrast, a view such as Artemov's (2008, section 6) which blocks Gettier cases by requiring knowledge-level justification to entail truth results in scepticism about induction.

3.4 Lottery cases and fallible justification

If I own a ticket in a fair lottery with one million tickets and only one winner, I have very good statistical grounds to believe that the ticket is a loser. Yet I do not seem to be in position to know that the ticket is a loser (Hawthorne, 2004, 1-2). On the method infallibilist approach, that is accounted for by saying that a method of forming the belief that the ticket is a loser on a merely statistical ground is a fallible one: if the lottery is fair, one world is accessible in which the ticket wins. Yet it yields true beliefs in *most* worlds, and that reflects the

[21] 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 a world is accessible in which the neighbour's habits are different. Or a world is accessible in which she believes something false on the same basis, for instance that some *other* neighbours are home while they are not. Knowledge would then require a method that is somehow sensitive to the neighbour's habits: for instance, one that would not lead my mother to 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 home tonight if their light is on tonight. Even in these more realistic settings, inductive knowledge boils down to *local* infallibility, infallibility in the kind of world one is in. (The account of inductive knowledge assumed here is externalist: whether a subject gains 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.)

fact that it provides a good statistical justification for the hypothesis that the ticket is a loser.[22]

These considerations allow us to draw an important distinction between two properties that are conflated under the label "fallibility". A method can be "fallible" in the sense that (a) it is fallible properly, that is, it produces false beliefs *within the relevant set of worlds*, (b) its infallibility is metaphysically contingent, that is, it produces false beliefs *outside* the relevant set of worlds. The distinction reflects an intuitive difference between Gettier cases and lottery cases. In typical Gettier cases, one's failure to know appears due to a peculiarity in one's situation. We account for that by saying that even though one's method is fallible, there are more normal worlds in which it would be infallible. By contrast, in lottery cases, one's method appears constitutively unable to deliver knowledge. We account for that by saying that one's method is fallible, if not at all worlds, at least at all the fairly normal ones.[23]

Properly fallible methods can be more or less reliable. 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 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 loser, we may say that the reliability of m is .99. This gives a sense in which m provides a good statistical support for the hypothesis that one's ticket will lose even though it falls short of providing knowledge.

A method that is reliable but properly fallible at a world is *merely reliable* at that world. Merely reliable methods yield fallibly justified beliefs. In section

[22] One may object there are very normal and even actual cases in which one's lottery belief is infallibly based. The lottery may be fixed, or one's ticket number may not have been included in the draw by accident. In such cases, the objection goes, one does not know and yet one's method is infallible. The method infallibilist's reply is that there has to be *other* actual or counterfactual lotteries in which one would form a belief on a basis that would count as the same, and that these cases are accessible from the present situation.

Conversely, one may object that in some cases, we gain knowledge on the basis of methods that can be described as merely statistical. Much scientific evidence is explicitly statistical, and on certain interpretations of quantum mechanics, any non-trivial belief about the future has some objective chance of being false (Hawthorne and Lasonen-Aarnio, 2009). To endorse such cases as cases of knowledge, the method infallibilist must say that either the error possibilities consistent with the statistical ground or the chances would not count as involving the same method, or that they are not accessible. See Williamson (2009) for the second strategy and Williams (2008) for a similar one in the context of semantics of counterfactuals.

[23] See Smith (2010, 10-11) for a similar view. There is no representation of the normality of worlds in our models. However, one could easily introduce it through a Lewisean system of spheres (Lewis, 1973) that orders worlds by relative normality, as Smith (2010) does to account for what he calls "fallible" justification. We leave the development to further work.

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.

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 model Gettier cases, inductive knowledge, Lottery cases and fallible justification. 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 simple models have something interesting to say. We provide a language and its semantics. Agents are specified as sets of methods and knowledge and belief are defined in terms of those. We state basic equivalence results with Scott-Montague neighbourhood models (Montague, 1968, 1970; Scott, 1970).

4.1 Propositions and the problem of Frege cases

We take propositions to be sets of worlds. Methods are thus functions from worlds and sets of sets of worlds to sets of sets of worlds. The resulting models are an extension of neighbourhood models, in which modality is represented by a relation from worlds to sets of sets of worlds (see appendix B).

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) p without knowing (or believing) q, where p and q are true at exactly the same worlds. [24] Well-known candidates involve logical equivalents, co-referential singular terms or names (Frege, 1892/1980, Kripke, 1979), indexicals and demonstratives (Kaplan, 1989, sec XVII, Perry, 1979) and natural kinds terms (Kripke, 1980). For instance, we are tempted to say that some Ancients knew that Hespherus was shone at night without knowing that Phosphorus did, even though every world in which Hespherus shone is one which Phosphorus did, and conversely, since they are one and the same planet, Venus. But if propositions are sets of worlds, the proposition that Hespherus shone is the same as the proposition that Phosphorus shone. So one cannot have a infallibly based belief in the first without having one in the second. Hence

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

methods models based on the coarse notion of proposition appear unadequate to represent referential opacity cases.

The matter is not so straightforward, however. We are tempted to ascribe a Phosphorus-belief when an Ancient uses a certain method — roughly, looking at the sky in the morning —, while we are tempted to ascribe a Hespherus-belief when they use another method — looking at the sky in the evening. Call them the Phosphorus-method and the Hespherus-method, respectively. We may thus describe the situation as follows: the Ancient believes that Venus shines on the basis of the Hespherus-method but not on the basis of the Phosphorus-method. The description may be used to explain away the intuition that "The Ancient does not believes that Phosphorus shines" is true: it is a way of expressing the fact that they do not believe it on the basis of the Phosphorus-method. The difference in method may also be taken to capture Frege's elusive notion of "mode of presentation": different modes of presentations correspond to different bases of belief.

The strategy has two limitations, however. First, referential opacity arises in cases that are not naturally described as involving two methods. For instance, if I can see the end of a ship by one window, and its other end by another, I may rationally come to believe that there are two distinct ships (Perry, 1979, 483). We may be tempted to say that I believe that that ship [pointing at one end] is identical to that ship [pointing at the same end] while not believing that that ship [pointing at the same end] is identical to that ship [pointing at the other end]. Yet it may appear *ad hoc* to count the two beliefs as involving distinct methods. Second, the strategy ascribes contradictory beliefs in such cases, and fails to distinguish irrational contradictory beliefs from rational ones.

Some further enrichment of the models appears needed for a full treatment of referential opacity. One option is to distinguish propositions that are true at the same possible worlds by introducing impossible worlds. Another is to take the outputs of methods to be sentence-like structures, for instance formulas of the language.[25] The latter option may implement a variety of philosophical views: a language-of-thought account of belief, a Fregean account of propositions, a view of belief as a ternary relation between subjects, modes of presentations (represented by sentences) and coarse propositions. We do not explore such developments here.

[25] A similar strategy appears in Awareness semantics (Fagin and Halpern, 1988) and Fitting's models for the Logic of Proofs (Fitting, 2005).

4.2 Syntax

Definition 1. Let $\mathsf{MC} = \{\mathsf{m}, \mathsf{n}, \ldots\}$ 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:

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

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

We make free use of the connectives $\rightarrow$, $\wedge$, $\leftrightarrow$, defined as usually.

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 Semantics

4.3.1 Methods, union and application

Given a set of worlds, we can define propositions, the space of possible methods and the operation of method union and application:

Definition 2. *Methods.* Given a set of worlds W:

$P = \mathcal{P}(W)$ is the set of propositions.

$MS = (W \times \mathcal{P}(P)) \rightarrow \mathcal{P}(P)$ is the space of possible methods: namely, all functions from worlds and sets of propositions (premises) to sets of propositions (conclusions).

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 *application of m to n* is the method $(m \circ n)$ given by $(m \circ n)(w, \pi) = m(w, n(w, \pi))$ for any w, π.

Remark 1. *Notation.* By convention, worlds are notated w, w', ..., propositions p, p', ..., q, q', ..., sets of propositions 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 W", "$\forall w \in W$" and "$\{w \in W|\ldots\}$", and similarly for any $p \in P$, $\pi \subseteq P$ and $m \in MS$. Note that we use sans-serif letters for object language constants

and operators ($\mathsf{p}, \mathsf{q}, \mathsf{m}, \mathsf{n}, \mathsf{K}, \mathsf{B}$) and serif letters for meta-language constants and variables (p, q, m, n, K, B).

When $\pi' = m(w, \pi)$, we say that π' is the *set of conclusions* reached by m at 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* reached by m at w. (See section 2.1.)

$m(w, \emptyset)$, the set of unconditionals outputs of m at w, is abbreviated $m(w)$.

$\langle MS, +, \circ \rangle$ is an algebraic structure over the space of possible methods. Its main properties are detailed in appendix A.

4.3.2 Frames and infallible methods

Our frames are given by a set of worlds W, an accessibility relation R over them, and a set of methods M^B for the agent.

Definition 3. A *methods frame* $\mathfrak{F}$ is a triple $\langle W, M^B, R \rangle$ where:
W is a non-empty set of worlds,
$M^B \subseteq MS$, with MS given by Definition 2,
$R \subseteq W \times W$ is a reflexive accessibility relation over worlds.

The background alethic modality $\langle W, R \rangle$ is used to define a set of infallible methods at each world. Recall that a method is infallible if and only if at all accessible worlds, it only derives true conclusions from sets of true premises (see section 2.3):

Definition 4. $M^I(w)$ is the set of *infallible methods at* w:
$m \in M^I(w)$ iff for all w', π: if wRw' and $w' \in p$ for all $p \in \pi$, then $w' \in q$ for all $q \in m(w', \pi)$.[26]

This implies in particular that all unconditional outputs are true at accessible worlds: if $m \in M^I(w)$, then for any w', p, if wRw' and $p \in m(w')$ then $q \in w'$. Note that infallibility is defined for all methods, irrespective of whether an agent has them or not.

Corollary 1. *Union and application 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. Suppose that $m \in M^I(w)$ and $n \in M^I(w)$, and let w' be any world such that wRw' and π be any set of propositions. By Definition 2, $q \in m+n(w', \pi)$ iff $q \in m(w', \pi)$ or $q \in n(w', \pi)$. Since m and n are infallible at w, by Definition 4, if $w' \in p$ for any $p \in \pi$, then $w' \in q$. So $m + n$ is infallible at w. Moreover,

[26] Equivalently, using the notion of information introduced in Appendix F: $m \in M^I(w)$ iff for any w' s.th. wRw': $\forall \pi(w' \in I(\pi) \rightarrow w' \in I(m(w', \pi)))$.

$q \in m \circ n(w', \pi)$ iff $q \in m(w, n(w', \pi))$. Since n is infallible at w, by Definition 4 again, if $w' \in p$ for any $p \in \pi$, then $w' \in q'$ for any $q' \in n(w, \pi)$. And since m is infallible at w, if $w' \in q'$ for any $q' \in n(w', \pi)$, then $w' \in q$ for any $q \in m(w, n(w'\pi))$. So $m \circ n$ is infallible at w. □

4.3.3 Agents, belief and knowledge

An agent is represented by a set of methods:

Definition 5. The agent's *method set,* notated M, is the *union and application closure of* M^B, that is the smallest set M such that $M^B \subseteq M$ and for any $m, n \in M$, $m + n \in M$ and $m \circ n \in M$.

When $m \in M$, we say that m *is one of the agent's methods.*

We assume an agent free of modular or computational limitations: the set of her methods is closed under application and union. Bounded agents can be modelled by putting restrictions on building M out of M^B (section 2.1).

Beliefs are just the unconditional outputs of the agent's methods. *Beliefs based of a certain method* are the unconditional outputs of that method, provided the agent has it. *Knowledge* is belief on an infallible basis. We define the functions corresponding to belief and knowledge on a basis and belief and knowledge *simpliciter* as follows ("B" and "K" each refer to functions of both types, but the ambiguity is convenient and their arguments always disambiguate):

Definition 6. *Belied and Knowledge.*

$B(m, w) = \{p | m \in M \wedge p \in m(w)\}$ gives the *agent's beliefs on the basis of* m *at* w. If m is not one of the agent's methods, $B(m, w)$ is empty for any w.

$K(m, w) = \{p | p \in B(m, w) \wedge m \in M^I(w)\}$ gives the *agent's knowledge on the basis of* m *at* w. If m is fallible at w, $K(m, w)$ is empty.

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

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

$B(w)$ 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.

The definitions of belief and knowledge *simpliciter* are not entirely innocuous. Our models allow agents whose methods unconditionally output both p and $\neg p$ at a world.[27] 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

[27] $\neg p$ stands for the negation of p, *i.e.* $W \backslash p$ on the coarse notion of proposition.

us from getting fully general theorems such as $\mathsf{Kp} \to \mathsf{KKp}$. That being said, I do not see the fact that we rely on this choice as an important liability. 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*.

4.3.4 Truth in a model

Definition 7. *Semantics.* Let $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ be a model where V is a function from propositional and methods constants to propositions and methods, respectively. 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.
$[\![\mathsf{p}]\!]^{\mathfrak{M}} = V(\mathsf{p})$,
$[\![\top]\!]^{\mathfrak{M}} = W$,
$[\![\neg\phi]\!]^{\mathfrak{M}} = W \backslash [\![\phi]\!]^{\mathfrak{M}}$,
$[\![\phi \vee \psi]\!]^{\mathfrak{M}} = [\![\phi]\!]^{\mathfrak{M}} \cup [\![\psi]\!]^{\mathfrak{M}}$.
Necessity, belief and knowledge.
$[\![\Box\phi]\!]^{\mathfrak{M}} = \{w : \forall w'(wRw' \to w' \in [\![\phi]\!]^{\mathfrak{M}})\}$,
$[\![\mathsf{B}\mu : \phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in B([\![\mu]\!]^{\mathfrak{M}}, w)\}$,
$[\![\mathsf{K}\mu : \phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in K([\![\mu]\!]^{\mathfrak{M}}, w)\}$,
$[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in B(w)\}$,
$[\![\mathsf{K}\phi]\!]^{\mathfrak{M}} = \{w : [\![\phi]\!]^{\mathfrak{M}} \in K(w)\}$.
Truth. $\models^{\mathfrak{M}}_{w} \phi$ iff $w \in [\![\phi]\!]^{\mathfrak{M}}$.
Validity. $\models^{\mathfrak{M}} \phi$ iff for any w, $\models^{\mathfrak{M}}_{w} \phi$.

When the intended model is clear from context, we write $[\![\cdot]\!]$ instead of $[\![\cdot]\!]^{\mathfrak{M}}$.

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 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 $\mathsf{K}\phi \to \phi$ is valid, there is an equivalent method model for K, and conversely (Theorems 21 and 22). 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. $\mathsf{B}\mu : \phi \to \mathsf{B}\phi$ *and* $\mathsf{K}\mu : \phi \to \mathsf{K}\phi$ *are valid in any methods model.*

Proof. By the semantics, if $\models^{\mathfrak{M}}_{w} \mathsf{B}\mu : \phi$ then there is an $m \in M$ such that $[\![\phi]\!] \in m(w)$. By Definition 6, $[\![\phi]\!] \in B(w)$, and by the semantics again, $\models^{\mathfrak{M}}_{w} \mathsf{B}\phi$. And analogously for K. □

5 Formal Results

The consequences of methods models fall within four groups:

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 our simpler models (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 axioms **K**: $\mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$ and **N**: $\mathsf{K}\top$.

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 the background alethic modality is transitive, *positive epistemic 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 *negative epistemic introspection* or axiom **5** ($\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$).

With *perfect 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 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 $[\![\phi]\!] = [\![\psi]\!]$, so $[\![\mathsf{B}\phi]\!] = \{w | [\![\phi]\!] \in B(w)\} = \{w | [\![\psi]\!] \in B(w)\} = [\![\mathsf{B}\psi]\!]$, and similarly for K. □

Referential transparency corresponds to the rule of equivalence of classical modal logic. It is known to determine the class of all neighbourhood frames (Chellas, 1980, 257). Thus by Theorem 21 (Appendix B) 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 limitation of the models, as noted above (section 4.1).

5.1.2 Factivity and subjectivity

Theorem 2. *Subjectivity (**S**).* $\mathsf{K}\phi \rightarrow \mathsf{B}\phi$ *is valid in any methods model.*

Proof. Evident from the semantics and Definitions 6. □

[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's (1996). Each has problematic consequences, as we noted in section 1. Thanks to an anonymous referee here.

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.[29]

Theorem 3. *Factivity ($\boldsymbol{T}_{\mathsf{K}}$).* $\mathsf{K}\phi \to \phi$ *is valid in any methods model.*

Proof. Evident from the fact that R is reflexive, the semantics, and the definitions of infallibility and knowledge (Definitions 4 and 6). □

Factivity entails that knowledge is consistent, or axiom **D**: $\mathsf{K}\phi \to \neg\mathsf{K}\neg\phi$.

It is known from neighbourhood semantics that $\mathbf{E}_{\mathsf{BK}}$ and $\mathbf{T}_{\mathsf{K}}$ determine truthful neighbourhood models.[30] By Theorem 22 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, $\Box$ should be a modality that itself satisfies **T** :

Theorem 4. *Alethic necessity ($\boldsymbol{T}_{\Box}$).* $\Box\phi \to \phi$ *is valid for any methods model.*

Proof. From the semantics and the reflexivity of R. □

This reflects the fact that $\Box$ is intended as an alethic modality. Without the reflexivity of R, a method may have all its unconditional outputs true at accessible worlds while having some false output in the actual world.

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 3.4). 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.

Without both conditions truth must be added as a *separate* condition on knowledge: p is known at w iff it is the unconditional output of some method "infallible" at w *and* p *is true at* w. 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. To avoid Gettier problems, a strict infallibility condition is needed (Sturgeon, 1993).

[29] 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 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".

[30] A neighbourhood model $\langle W, N\rangle$ is truthful iff $w \in \bigcap N(w)$ for any w, where $\bigcap N(w)$ is set to W when $N(w)$ is empty. See Appendix B.

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, and* ***D*** *and* ***T*** *for* B *fail in some methods model.*

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

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

$\boldsymbol{M}_{\mathsf{B}}$. $\mathsf{B}(\phi \wedge \psi) \to (\mathsf{B}\phi \wedge \mathsf{B}\psi)$

$\boldsymbol{C}_{\mathsf{B}}$. $(\mathsf{B}\phi \wedge \mathsf{B}\psi) \to \mathsf{B}(\phi \wedge \psi)$

$\boldsymbol{K}_{\mathsf{B}}$. $\mathsf{B}(\phi \to \psi) \to (\mathsf{B}\phi \to \mathsf{B}\psi)$

$\boldsymbol{N}_{\mathsf{B}}$. $\mathsf{B}\top$

$\boldsymbol{4}_{\mathsf{B}}$. $\mathsf{B}\phi \to \mathsf{B}\mathsf{B}\phi$

$\boldsymbol{5}_{\mathsf{B}}$. $\neg\mathsf{B}\phi \to \mathsf{B}\neg\mathsf{B}\phi$

$\boldsymbol{M}_{\mathsf{K}}$. $\mathsf{K}(\phi \wedge \psi) \to (\mathsf{K}\phi \wedge \mathsf{K}\psi)$

$\boldsymbol{C}_{\mathsf{K}}$. $(\mathsf{K}\phi \wedge \mathsf{K}\psi) \to \mathsf{K}(\phi \wedge \psi)$

$\boldsymbol{K}_{\mathsf{K}}$. $\mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$

$\boldsymbol{N}_{\mathsf{K}}$. $\mathsf{K}\top$

$\boldsymbol{4}_{\mathsf{K}}$. $\mathsf{K}\phi \to \mathsf{K}\mathsf{K}\phi$

$\boldsymbol{5}_{\mathsf{K}}$. $\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$

Proof. Neighbourhood models that invalidate each schemas are known. By Theorem 21 they can be used to show that none of the B schema are valid in methods models. Moreover, for other schemas that **D** and **T**, it is easy to construct such counter-models as truthful neighbourhood models, so that by theorem 22 we have counterexamples to the K schemas. □

Appendix C gives examples of violations of **M**, **N**, **K**, and **4**. In particular, the violation of **K** illustrates our *Watson* case (section 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 two specific methods.

Definition 8. A *perfect reasoner model* is a methods model such that $m^R, m^D \in M$, where:

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

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.

We introduce constants m^{R} and m^{D} in $\mathcal{L}$ such that for any model $\mathfrak{M}$, $[\![\mathsf{m}^{\mathsf{R}}]\!]^{\mathfrak{M}} = m^R$ and $[\![\mathsf{m}^{\mathsf{D}}]\!]^{\mathfrak{M}} = m^D$.

Pure Reason outputs the tautology given any set of premises, including the empty set.[31] Deduction maps a set of premises to all logical consequences of

[31] 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 $w, \pi \neq \emptyset$.

any pair of premises.[32] A perfect reasoner is an agent that has Pure Reason and Deduction among its methods.[33]

Theorem 6. *Knowledge of logic ($\boldsymbol{N}_{\mathsf{B}}$ and $\boldsymbol{N}_{\mathsf{K}}$).* $\mathsf{Bm}^{\mathsf{R}} : \top$ *and* $\mathsf{Km}^{\mathsf{R}} : \top$ *are valid in any perfect reasoner model. By Corollary 2,* $\mathsf{B}\top$ *and* $\mathsf{K}\top$ *are valid as well.*

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 the definition of Pure Reason, , $W \in m^R(w)$. Since $m^R \in M$, by the definition of belief, $W \in B(w)$ (Definition 6). By the semantics, $\models^{\mathfrak{M}}_w \mathsf{B}\top$.

By the definition of Pure Reason again, for any p, w', π, if $p \in m^R(w', \pi)$ then $p = W$, so $w' \in p$. Thus by the definition of infallibility, $m^R \in M^I(w)$ (Definition 4). Since $m^R \in M$, by the definition of knowledge (Definition 6), $W \in K(w)$, and by the semantics, $\models^{\mathfrak{M}}_w \mathsf{K}\top$. □

Theorem 7. *Logical omniscience.* $\mathsf{B}\mu : (\phi \to \psi) \to (\mathsf{B}\nu : \phi \to \mathsf{Bm}^{\mathsf{D}} \circ (\mu + \nu) : \psi)$ *and* $\mathsf{K}\mu : (\phi \to \psi) \to (\mathsf{K}\nu : \phi \to \mathsf{Km}^{\mathsf{D}} \circ (\mu + \nu) : \psi)$ *are valid in any perfect reasoner model. By Corollary 2, ($\boldsymbol{K}_{\mathsf{B}}$)* $\mathsf{B}(\phi \to \psi) \to (\mathsf{B}\phi \to \mathsf{B}\psi)$ *and ($\boldsymbol{K}_{\mathsf{K}}$)* $\mathsf{K}(\phi \to \psi) \to (\mathsf{K}\phi \to \mathsf{K}\psi)$ *are valid as well.*

Proof. We show that whenever $p \to q$ and p are believed on the basis of m and n respectively, q is believed by $m^D \circ (m + n)$. 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 the semantics and the definition of belief (Definition 6), $(W \backslash [\![\phi]\!]) \cup [\![\psi]\!] \in [\![\mu]\!](w)$, $[\![\phi]\!] \in [\![\nu]\!](w)$ and $[\![\mu]\!], [\![\nu]\!] \in M$. By the definition of method union, $(W \backslash [\![\phi]\!]) \cup [\![\psi]\!], [\![\phi]\!] \in ([\![\mu]\!] + [\![\nu]\!])(w)$. Since $((W \backslash [\![\phi]\!]) \cup [\![\psi]\!]) \cap [\![\phi]\!] \subseteq [\![\psi]\!]$, by the definitions of application and Multi-Premise Deduction, $[\![\psi]\!] \in m^D \circ ([\![\mu]\!] + [\![\nu]\!])(w)$. Since the agent is a perfect reasoner, $m^D \in M$, and since the method set is closed under union and application,

Conditional Pure Reason outputs the tautology only if the agent has some other belief. The resulting schema are $\mathsf{B}\phi \to \mathsf{B}\top$ and $\mathsf{B}\phi \to \mathsf{K}\top$ for any ϕ.

[32] Note that it outputs nothing of the basis of the empty set of premises. Given any premise, Deduction outputs the tautology. But that does not make Pure Reason redundant. If all non-inferential methods of the agent are fallible, then applying Deduction to them cannot yield knowledge, since the resulting 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.

[33] Note that an agent may be a perfect reasoner without having Pure Reason and Deduction in its *basic* set M^B. To illustrate, consider M^B such that $m^R \notin M^B$ but $m, n \in M^B$ where m outputs the tautology at p worlds and n outputs the tautology at $\neg p$-worlds (for some arbitrary proposition p), *i.e.* $m(w, \pi) = \{W\}$ if $w \in p$ and $\emptyset$ otherwise, and $n(w, \pi) = \{W\}$ if $w \notin p$ and $\emptyset$ otherwise. We have $m + n = m^R$. Since $m, n \in M^B$ and M is the union and application closure of M^B, $m + n = m^R \in M$, even though $m^R \notin M^B$. So the agent is a perfect reasoner without having Pure Reason among her basic methods.

$m^D \circ (\llbracket\mu\rrbracket + \llbracket\nu\rrbracket) \in M$ (Definitions 8 and 5). By Definition 6 again, $\llbracket\psi\rrbracket \in B(m^D \circ (\llbracket\mu\rrbracket + \llbracket\nu\rrbracket), w)$, and by the semantics, $\models^{\mathfrak{M}}_w \mathsf{Bm}^{\mathsf{D}} \circ (\mu + \nu) : \psi$, which completes the proof of $\mathbf{K}_{\mathsf{B}}$.

For infallibility, let $\mathfrak{M}, w$ be any perfect reasoner model and world. Let w' be such that wRw' and π be such that $w' \in p$ for any $p \in \pi$. Suppose $q \in m^D(w', \pi)$. By the definition of Multi-Premise Deduction, there are $p', p'' \in \pi$ such that $q \supseteq p' \cap p''$. Since $w' \in p'$, $w' \in p''$ and $q \supseteq p' \cap p''$, $w' \in q$. Generalising over w', π, m^D is infallible at w (Definition 4).

Now suppose $\models^{\mathfrak{M}}_w \mathsf{K}\mu : (\phi \to \psi) \wedge \mathsf{K}\nu : \phi$ for a perfect reasoner model $\mathfrak{M}$, a world w, and some ϕ,ψ. The situation is as before with, additionally, $\llbracket\mu\rrbracket, \llbracket\nu\rrbracket \in M^I(w)$, by the semantics and the definition of knowledge (Definition 6). Since $m^D \in M^I(w)$ and since union and application preserve infallibility (Corollary 1), $m^D \circ (\llbracket\mu\rrbracket + \llbracket\mu\rrbracket) \in M^I(w)$. We show as before that $\llbracket\psi\rrbracket \in m^D \circ (\llbracket\mu\rrbracket + \llbracket\mu\rrbracket)(w)$ and $m^D \circ (\llbracket\mu\rrbracket + \llbracket\mu\rrbracket) \in M$, so by the semantics, $\models^{\mathfrak{M}}_w \mathsf{Km}^{\mathsf{D}} \circ (\mu + \nu) : \psi$, 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 is captured by the following theorem:

Theorem 8. *Deductive closure.* $\mathsf{K}\mu : \phi \to (\mathsf{Bm}^{\mathsf{D}} \circ \mu : \psi \to \mathsf{Km}^{\mathsf{D}} \circ \mu : \psi)$ *is valid in any methods model. (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 μ, ϕ. By the definition of knowledge, $[\![\mu]\!] \in M^I(w)$ (Definition 6). Since Deduction is infallible (see Theorem 7) and application preserves infallibility (Corollary 1), $m^D \circ [\![\mu]\!] \in M^I(w)$. Now suppose in addition that $\models^{\mathfrak{M}}_w \mathsf{Bm}^{\mathsf{D}} \circ \mu : \psi$ for some ψ. Since $[\![\mathsf{m}^{\mathsf{D}} \circ \mu]\!] = m^D \circ [\![\mu]\!] \in M^I(w)$, we have $\models^{\mathfrak{M}}_w \mathsf{Km}^{\mathsf{D}} \circ \mu : \psi$. □

The theorem holds even if the agent's method set is not closed under union and application, and it has no equivalent in a language without methods terms.[34]

Further exploration of the deductive aspects of methods models are made in Appendix D.

5.3 Consistency

Pure Reason and Deduction do not guarantee than an agent is consistent: that is, the $\mathbf{D}_B$ axiom for belief ($\mathsf{B}\phi \to \neg\mathsf{B}\neg\phi$) may fail. (The $\mathbf{D}_K$ axiom for knowledge is of course guaranteed by the factivity of knowledge, see 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 9. A *consistent agent model* is a methods model such that for any w, $\bigcap B(w) \neq \emptyset$, where $\bigcap B(w) = W$ whenever $B(w)$ is empty.

Theorem 9. *($\boldsymbol{D}_B$) $\mathsf{B}\phi \to \neg\mathsf{B}\neg\phi$ is valid for any consistent agent model.*

Proof. Let $\mathfrak{M}, w$ be a consistent agent model and a world such that $\models^{\mathfrak{M}}_w \mathsf{B}\phi$ for some ϕ. By the definition of belief, $[\![\phi]\!] \in B(w)$. Since the agent is consistent, $W \backslash [\![\phi]\!] \notin B(w)$. So $\models^{\mathfrak{M}}_w \neg\mathsf{B}\neg\phi$. □

5.4 Perfect introspection and perfect confidence

The next two interesting classes of models are those of *perfect introspecters* and *confident introspecters*. In intuitive terms, 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

[34] The closest we can formulate without methods terms is $\mathsf{K}\phi \to ((\Box(\phi \to \psi) \wedge \mathsf{B}\psi) \to \mathsf{K}\psi)$, which is counter-exampled if the agent believes ψ from some other reasons than ϕ, as in our *Watson* case. $\mathsf{K}\phi \to (((\phi \to \psi) \wedge \mathsf{B}\psi) \to \mathsf{K}\psi)$ would reduce to $\mathsf{K}\phi \to ((\psi \wedge \mathsf{B}\psi) \to \mathsf{K}\psi)$, which, barring bizarre cases, holds only for excellent agents: see Definition 13, section 5.5.

knows it, and whenever she does not believe something, she believes that she does not know it.

By contrast with Reason and Deduction, Introspection and Confident Introspection methods cannot be characterised in a model-independent manner. As we characterised them, introspection methods outputs propositions of the kind: *the agent believes that p*. But since we model propositions as sets of worlds, "the" proposition that *the agent believes that p* is actually a different proposition from one model to another: even with p kept constant, the set of worlds at which the agent believes that p may vary from a model to another, depending on what methods the agent has. For the same reason, in multi-agent settings, introspection and confidence methods should be further relativised to agents. The relativisation of these methods to model and agent reflects the fact that these methods tell us something about the agent's states.[35]

We notate propositions about what the agent believes and knows as follows:

Definition 10. For any methods model $\mathfrak{M}$:

$b(\mathfrak{M}, p) = \{w | p \in B(w)\}$ is, in $\mathfrak{M}$, the proposition that the agent believes that p,

$-b(\mathfrak{M}, p) = \{w | p \notin B(w)\} = W \backslash b(\mathfrak{M}, p)$ is its negation, and analogously for knowledge:

$k(\mathfrak{M}, p) = \{w | p \in K(w)\}$,

$-k(\mathfrak{M}, p) = \{w | p \notin K(w)\} = W \backslash k(\mathfrak{M}, p)$.

It is easy to check that for any $\mathfrak{M}, \phi$, $[\![\mathsf{B}\phi]\!]^{\mathfrak{M}} = b(\mathfrak{M}, [\![\phi]\!]^{\mathfrak{M}})$, $[\![\neg\mathsf{B}\phi]\!]^{\mathfrak{M}} = -b(\mathfrak{M}, [\![\phi]\!]^{\mathfrak{M}})$, $[\![\mathsf{K}\phi]\!]^{\mathfrak{M}} = k(\mathfrak{M}, [\![\phi]\!]^{\mathfrak{M}})$, $[\![\neg\mathsf{K}\phi]\!]^{\mathfrak{M}} = -k(\mathfrak{M}, [\![\phi]\!]^{\mathfrak{M}})$ (semantics and Definition 6).

We omit reference to the model when it is clear from the context, and simply write $b(p)$, $-b(p)$, $k(p)$, $-k(p)$.

5.4.1 Perfect introspection: self-knowledge

Definition 11. A *perfect introspecter model* is a methods model $\mathfrak{M}$ such that $pi(\mathfrak{M}, M), ni(\mathfrak{M}, M) \in M$ where:

Given any methods model $\mathfrak{N}$ and any set of methods X, the *Positive Introspection of X in* $\mathfrak{N}$, notated $pi(\mathfrak{N}, X)$, is the method such that for any w, p, π:

[35] A contrast may be useful here. 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, $!m$ 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.

$p \in pi(\mathfrak{N}, X)(w, \pi)$ iff for some p', $p = b(\mathfrak{N}, p')$ and there is some $m \in X$ such that $p' \in m(w)$.

Given any methods models $\mathfrak{N}$, the *Negative Introspection of* X *in* $\mathfrak{N}$, notated $ni(X)$, is the method such that for any w, p, π: $p \in ni(\mathfrak{N}, X)(w, \pi)$ iff for some p', $p = -b(\mathfrak{N}, p')$ and there is no $m \in X$ such that $p' \in m(w)$.

We omit reference to the model when possible and simply write $pi(X)$ and $ni(X)$.

For each method m in X, the method $pi(X)$ outputs that the agent believes p whenever m outputs p. In intuitive terms, an agent that has the method $pi(X)$ takes herself to have at least the methods in X, as far as her beliefs are concerned. Similarly, the method $ni(X)$ outputs that the agent does not believe p, whenever 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. But note that these Introspection methods are defined for any model and any set of methods: neither the methods in X nor $pi(X)$ and $ni(X)$ need be among the agent's methods.[36]

It is easy to see that $pi(X)$ is infallible if the agent's methods include those in X and that $ni(X)$ is infallible if the agent's methods are included in X:

Lemma 1. *For any methods model:*

For any $X \subseteq M$, the Positive Introspection of X is infallible.

For any $X \supseteq M$, the Negative Introspection of X is infallible.

Proof. Positive introspection. Let $\mathfrak{M}, X$ be such that $X \subseteq M$ and let w be any world. Suppose that wRw' and $p \in pi(X)(w', \pi)$ for some w', p, π. By the definition of Positive Introspection, there is some m, p' such that $p = b(p')$, $m \in X$ and $p' \in m(w')$. Since $X \subseteq M$, $m \in M$. Since $p' \in m(w')$ and $m \in M$, by the definition of belief, $p' \in B(w')$ (Definition 6). Since $p' \in B(w)$, $w' \in b(p')$, *i.e.* $w' \in p$. Generalising over w', p, $pi(X)$ is infallible at w (Definition 4).

Negative introspection. Let $\mathfrak{M}, X$ be such that $M \subseteq X$ and let w be any world. Suppose that wRw' and $p \in ni(X)(w', \pi)$ for some w', p, π. By the definition of Negative Introspection, 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 the definition of belief, $p \notin B(w')$. Hence $w' \in -b(p')$, *i.e.* $w' \in p$. Generalising over w', p, $ni(X)$ is infallible at w. □

Perfect introspecters have positive and negative introspection methods that perfectly match her set of methods: $pi(M)$ and $ni(M)$. Perfect introspecters have perfect knowledge of their own beliefs:

[36] Our introspection methods only reflect the *unconditional* outputs of an agent's methods, *i.e.* her beliefs. One could imagine richer introspection methods that also reflect an agent's inferential dispositions or conditional beliefs.

Theorem 10. *Self-knowledge (**SK**).* $\mathsf{B}\phi \rightarrow \mathsf{KB}\phi$ *and* $\neg\mathsf{B}\phi \rightarrow \mathsf{K}\neg\mathsf{B}\phi$ *are valid for any perfect introspecter model.*

Proof. Positive self-knowledge. Suppose $\mathfrak{M}$ is a perfect introspecter model and w any world such that $\models^{\mathfrak{M}}_{w} \mathsf{B}\phi$ for some ϕ. By the definition of belief, there is some $m \in M$ such that $[\![\phi]\!] \in m(w)$. By the definition of positive introspection, $b([\![\phi]\!]) \in pi(M)(w)$, where $b([\![\phi]\!]) = [\![\mathsf{B}\phi]\!]$ (Definition 10). Since the agent is a perfect introspecter, $pi(M) \in M$. By Lemma 1, $pi(M) \in M^I(w)$. So by the definition of knowledge, $\models^{\mathfrak{M}}_{w} \mathsf{KB}\phi$.

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 the definition of belief, there is no $m \in M$ such that $[\![\phi]\!] \in m(w)$. By the definition of negative introspection, $-b([\![\phi]\!]) = [\![\neg\mathsf{B}\phi]\!] \in ni(M)(w)$. Since the agent is a perfect introspecter, $ni(M) \in M$, and by Lemma 1, $ni(M) \in M^I(w)$. By the definition of knowledge, $\models^{\mathfrak{M}}_{w} \mathsf{K}\neg\mathsf{B}\phi$. □

Our perfect Positive Introspection method is coarse-grained: one method for all of the agent's beliefs. We could have used more fine-grained methods, *e.g.* one per method: the set of $pi(\{m\})$ where $m \in M$. By contrast, perfect Negative Introspection is essentially holistic: we cannot get it on a method-per-method basis. Consider $ni(\{m\})$ for some $m \in M$: the method outputs that the agent does not believe p whenever *the method* m does not output p. The method $ni(\{m\})$ will go wrong in cases in which the agent believes that p on the basis of some other method than m. At most we can define fine-grained infallible methods that output that the agent does not believe p *on the basis of* m.[37] But negative introspection methods of this type cannot deliver the proposition that the agent does not believe p *simpliciter*.[38] This asymmetry between Positive and Negative Introspection is due to the fact that it is sufficient for an agent to believe p that *some* of her method outputs p, while for her *not* to believe p, it is necessary that *none* of her methods outputs p.

Perfect introspection methods are further discussed in Appendix E.

5.4.2 Perfect Confidence: partial knowledge of one's ignorance and knowledge of one's knowledge

Confident Introspection methods are introspection methods whose output is the proposition that the agent knows instead of the proposition that the agent believes:

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

[38] 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.

Definition 12. A *(perfect) confident introspecter model* is a methods model such that for each $m \in M$, $pc(\{m\}) \in M$, and $nc(M) \in M$, where:

Given any methods model $\mathfrak{N}$ and method set X, the *Positive Confident Introspection of X in $\mathfrak{N}$*, noted $pc(\mathfrak{N}, X)$, is the method such that for any w, p, π: $p \in pc(\mathfrak{N}, X)(w, \pi)$ iff for some p', $p = k(\mathfrak{N}, p')$ and there is a $m \in X$ such that $p' \in m(w)$.

Given any methods model $\mathfrak{N}$ and method set X, the *Negative Confident Introspection of X*, noted $nc(\mathfrak{N}, X)$, is the method such that for any w, p, π: $p \in nc(\mathfrak{N}, X)(w, \pi)$ iff for some p', $p = -k(\mathfrak{N}, p')$ and there is no $m \in X$ such that $p' \in m(w)$.

As before, we omit reference to models when it is clear from the context.

In intuitive terms, an agent that has Positive Confidence of X believes as if all the methods in X gave her knowledge, and an agent that has Negative Confidence 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 2. *For any methods model:*

For any $X \supseteq M$, the Negative Confident Introspection of X is infallible.

Proof. We transpose the proof of Lemma 1, using the fact that if $p' \notin B(w')$ then $p' \notin K(w')$ (Definition 6 and Theorem 2) . □

For Positive Confidence we would expect a *conditional* infallibility result: if all the methods in X are infallible methods of the agent, then $pc(X)$ is infallible. However, this can only be proved if the background accessibility relation is transitive — that is, if what is possibly possible is possible. For any output of a method in X, $pc(X)$ tells that the agent knows it. If the methods in X are infallible and are among the agent's methods, that means that all the outputs of $pc(X)$ are true. But for $pc(X)$ to be *infallible*, its outputs must not only be true at the actual world, but also at accessible worlds. So the methods in X must be infallible at accessible worlds as well. Now if accessibility is transitive, the infallibility of a method at a world guarantees its infallibility at accessible worlds:

Lemma 3. *If $\mathfrak{M}$ is a methods model with a transitive R, then for any m, w, w' such that wRw', if $m \in M^I(w)$ then $m \in M^I(w')$.*

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 the definition of infallibility, $\forall\pi((\forall p \in \pi(w'' \in p)) \rightarrow \forall q \in m(w'', \pi)(w'' \in q))$. Generalising over w'', by the definition of infallibility again, $m \in M^I(w')$. □

This gives us the desired result:

Lemma 4. *For any transitive methods models:*

For any $X \subseteq M$, if $X \subseteq M^I(w)$ then the Positive Confident Introspection of X is infallible.

Proof. Let $\mathfrak{M}$ be a model with transitive R, X such that $X \subseteq M$ and w a world such that $X \subseteq M^I(w)$. Suppose that wRw' and $p \in pc(X)(w', \pi)$ for some w', p, π. By the definition of Confident Introspection, there is some p' such that $p = k(p')$ and $p' \in m(w')$ for some $m \in X$. Since R is transitive, $m \in M^I(w)$ and wRw', $m \in M^I(w')$ (Lemma 3). Since $X \subseteq M$, $m \in M$. Since $p' \in m(w')$ and $m \in M^I(w')$, by the definition of knowledge, $p' \in K(w')$. Hence $w' \in k(p')$. Generalising over w', p, π, $pc(X)$ is infallible. □

An agent is a Perfect Confident Introspecter if she has Confident Introspection methods that exactly match her method set: $pc(\{m\})$ for each $m \in M$, and $nc(M)$. Note that our Positive Confidence Introspection methods are very fine-grained: one per method. A coarse-grained $pc(M)$ method would be fallible as long as the agent has one fallible method. Alternative options are discussed in Appendix E.

We now establish three results.

Theorem 11. *Confident Introspection.* $\mathsf{B}\phi \rightarrow \mathsf{BK}\phi$ *and* $\neg\mathsf{B}\phi \rightarrow \mathsf{B}\neg\mathsf{K}\phi$ *are valid in confident introspecter models.*

Proof. Evident from the semantics and the definitions of belief, knowledge and confident introspecter. □

Theorem 12. *Partial knowledge of one's ignorance (**p5**).* $\neg\mathsf{B}\phi \rightarrow \mathsf{K}\neg\mathsf{K}\phi$ *is valid in confident introspecter models.*

Proof. We transpose the proof of negative self-knowledge (Theorem 10), using the infallibility of Negative Confidence (Lemma 2). □

Given subjectivity ($\mathsf{K}\phi \rightarrow \mathsf{B}\phi$), the theorem is equivalent to a conditionalised version of axiom **5**: $\neg\mathsf{B}\phi \rightarrow (\neg\mathsf{K}\phi \rightarrow \mathsf{K}\neg\mathsf{K}\phi)$. I am using "ignorance" here in a slightly unnatural way for the negation of knowledge. (Thus if p is false p is part of the subject's "ignorance".)

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 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 find out 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.

Theorem 13. *Knowledge of one's knowledge (4).* $\mathsf{K}\phi \rightarrow \mathsf{KK}\phi$ *is valid in confident introspective models with a transitive accessibility relation.*

Proof. Let $\mathfrak{M}$ be a perfect confident introspecter model in a transitive frame such that $\models^{\mathfrak{M}}_{w} \mathsf{K}\phi$ for some w, ϕ. By the definition of knowledge, there is an $m \in M$ such that $[\![\phi]\!] \in m(w)$ and $m \in M^I(w)$. By the definition of Positive Confidence, $k([\![\phi]\!]) = [\![\mathsf{K}\phi]\!] \in pc(\{m\})$. Since the agent is a confident introspecter, $pc(\{m\}) \in M$. Since the frame is transitive, and since $m \in M$ and $m \in M^I(w)$, by Lemma 4, $pc(\{m\}) \in M^I(w)$. By the definition of knowledge, $\models^{\mathfrak{M}}_{w} \mathsf{KK}\phi$. □

That is the only result for which we make a stronger assumption on the background accessibility relation than reflexivity. Transitivity is a natural assumption on some notions of alethic possibility, such as physical possibility, but not on others, such as closeness or similarity. On the latter view of the possibilities of error relevant to knowledge, epistemic introspection may easily fail (Williamson, 2000, chap. 5). It may appear surprising that **4** requires an additional assumption about possibility while **p5** does not. The point is discussed in appendix E.

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 13. A *excellent agent model* is a methods model such that $M \subseteq M^I(w)$ for any w.

Theorem 14. *Belief is knowledge (**BK**).* $\mathsf{B}\phi \leftrightarrow \mathsf{K}\phi$ *is valid for excellent agents.*

Proof. The right-to-left direction follows from Subjectivity. The left-to-right direction is evident from the semantics and the definitions of excellence, belief and knowledge. Let $\mathfrak{M}, w$ be an excellent agent model and world such that $\models^{\mathfrak{M}}_{w} \mathsf{B}\phi$ for some ϕ. By the semantics and the definition of belief, there is an $m \in M$ such that $[\![\phi]\!] \in m(w)$. Since the agent is excellent, $m \in M^I(w)$. By the definition of knowledge, $\models^{\mathfrak{M}}_{w} \mathsf{K}\phi$. □

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 knowledge (**4**) and Perfect knowledge of one's ignorance (**5**).* $\mathsf{K}\phi \to \mathsf{KK}\phi$ *and* $\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$ *are valid for excellent confident introspecter models.*

Proof. Suppose $\mathfrak{M}$ is an excellent agent model and m, w a method and a world such that $m \in M^I(w)$ and $m \in M$. Then by the definition of excellence, $m \in M^I(w')$ for any w'. In particular, $m \in M^I(w')$ for any w' such that wRw'. Thus we get the equivalent of Lemma 3 without assuming transitivity. From this $\models^{\mathfrak{M}}_w \mathsf{K}\phi \to \mathsf{KK}\phi$ is derived as in Theorem 13.

Suppose $\mathfrak{M}, w$ are an excellent confident introspecter model and a world such that $\models^{\mathfrak{M}}_w \neg\mathsf{K}\phi$. Since the agent is excellent, by Theorem 14, $\models^{\mathfrak{M}}_w \neg\mathsf{B}\phi$. Since the agent is a confident introspecter, by Theorem 12, $\models^{\mathfrak{M}}_w \mathsf{K}\neg\mathsf{K}\phi$. □

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 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 or draw false inferences. The simple **S5** epistemic system is suited to that use.

We can get a similar result with Introspection only, but it heavily relies on the fact that propositions are coarsely individuated:

Theorem 16. $\mathsf{K}\phi \to \mathsf{KK}\phi$ *and* $\neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$ *are valid for excellent perfect introspecter models.*

Proof. Let $\mathfrak{M}$ be a excellent perfect introspecter model. Since the agent is a perfect introspecter, $\models^{\mathfrak{M}} \mathsf{B}\phi \to \mathsf{KB}\phi$ and $\models^{\mathfrak{M}} \neg\mathsf{B}\phi \to \mathsf{K}\neg\mathsf{B}\phi$ (Theorem 10). Since the agent is excellent, $[\![\mathsf{B}\phi]\!] = [\![\mathsf{K}\phi]\!]$ and $[\![\neg\mathsf{B}\phi]\!] = [\![\neg\mathsf{K}\phi]\!]$ (Theorem 14). By referential transparency, $\models^{\mathfrak{M}} \mathsf{K}\phi \to \mathsf{KK}\phi$ and $\models^{\mathfrak{M}} \neg\mathsf{K}\phi \to \mathsf{K}\neg\mathsf{K}\phi$ (Theorem 1). □

The proof relies on the fact that in the case of excellent agents, the proposition that the agent believes something and the proposition that she knows something are one and the same. Thus perfect knowledge of one's beliefs amounts to perfect knowledge of one's knowledge, and the distinction between Introspection and Confidence methods is lost. For similar reasons, it is impossible to model a *modest* excellent agent who believes that she believes something without believing that she knows it. That is counter-intuitive, since such an agent may obviously exist. These facts reflect a limitation of our coarse models and should fail on finer-grained notions of propositions.[39]

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 unbounded 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 (b) and (c) together. 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 an agent believes and knows. (That is, if a possible world is incompatible with the agent's beliefs based on $m^D \circ m$, that possible world is also incompatible

[39] Thanks to Timothy Williamson here.

some belief of the agent based on m; the same does not hold for $pi(\{m\})$ and $pc(\{m\})$.) See Appendix F 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 thus provides guidance 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 philosophical epistemology, such as reliabilism, 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 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, in the case of positive epistemic introspection, 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. 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 sketch in appendix A 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). The latter would require us 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, our Confidence methods are *maximally immodest* and *non-inferential*. An inferential conception of Confidence is more natural, and corresponding methods should be defined. We should also consider modest agents whose epistemic confidence extends only to a subset of their methods. Third, our methods can be used to represent *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 hypotheses that π. Consequences of the definition should be explored, as well as extension of Introspection to conditional beliefs. Fourth, the idea of a *reliability measure* over methods that we have sketched in section 3.4 should be investigated in order to see whether it can yield an interesting notion of epistemic probability. The idea would be that the epistemic probability of a based belief is the probability that a belief with that basis is true. Some beliefs in logical truths, for instance, could receive an epistemic probability lesser than one, provided that they are produced by fallible methods.

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. Since our Introspection methods are not proprietary to the agent they introspect, they can be recast as *mind-reading* methods. They can then be used to define notions of common belief and knowledge. We sketch this development in Appendix E. However, the development does not take into account the idea that a same method may yield different outputs in different agents, i.e. *perspective-relativity* (see 2.1). Further work should integrate the idea.

Finally, the methods approach need not be restricted to epistemology. Along-

side 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 epistemological 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 the 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.

A Algebra for methods

$\langle \mathbf{M}, +, \circ, 0, 1 \rangle$ is an algebraic structure over the set of methods, where:

Definition 14. 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.[41]

Theorem 17. *Method union is idempotent, commutative and associative. For any methods m, n, r: $m+m = m$, $m+n = n+m$, and $(m+n)+r = m+(n+r)$.*

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

Remark 2. The *empty method* 0 is uniquely characterised as the method such that $0 + n = n$ for any n.

Theorem 18. *Method application is associative but not idempotent nor commutative. For any methods m, n, r: $m \circ (n \circ r) = (m \circ n) \circ r$. 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 $\neg(p \wedge q) \in (m \circ n)(w, \{p, q\})$ but $\neg(p \wedge q) \notin (n \circ m)(w, \{p, q\})$ at any w, so $m \circ n \neq n \circ m$. □

Remark 3. The identity method is uniquely characterised as the method such that $1 \circ n = n \circ 1 = n$ for any method m.

Remark 4. Non-inferential methods are methods which 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.

Remark 5. The empty method is non-inferential: $0 \circ n = 0$ for any n.

Remark 6. Applying a method m to the empty one amounts to applying it without premise: $(m \circ 0)(w, \pi) = m(w)$. Thus $m \circ 0$ is non-inferential for any 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 7. Call a method *purely conditional* iff it does not yield unconditional outputs. Purely conditional methods are characterised as the set of methods m such that $m \circ 0 = 0$.

Remark 8. The empty method and the identity methods are purely conditional.

Theorem 19. *Union does not distribute over application.* $m + (n \circ r) = (m + n) \circ (m + r)$ *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 20. *Application distributes right-to-left over union, but not left-to-right. For any methods* m, n, r*:* $(m + n) \circ r = (m \circ r) + (n \circ r)$*. 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). In particular, $(m + n)(w, r(w, \pi')) = m(w, r(w, \pi')) \cup n(w, r(w, \pi'))$. Thus by Definition 2, $(m + n) \circ r = m \circ r + n \circ r$.

Counterexample to left-to-right distribution. Write $p \vee q$ for 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 $n(w) = \{p\}$ and $r(w) = \{q\}$. Assuming $p \neq q$, we have $p \vee q \in m \circ (n + r)(w)$ but $p \vee q \notin (m \circ n) + (m \circ r)(w)$. □

Application distributes right-to-left but not left-to-right because the methods algebra represents information flow or informational dependencies. Applying m to $n + r$ 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 to 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.

Together with the semantics, the algebra gives us a range of equivalences:

Corollary 3. *The following are valid in any methods model:*

$\mathsf{B}(\mu + \nu) + \rho : \phi \leftrightarrow \mathsf{B}\mu + (\nu + \rho) : \phi$,
$\mathsf{B}\mu + \nu : \phi \leftrightarrow \mathsf{B}\nu + \mu : \phi$,
$\mathsf{B}\mu + \mu : \phi \leftrightarrow \mathsf{B}\mu : \phi$,
$\mathsf{B}(\mu \circ \nu) \circ \rho : \phi \leftrightarrow \mathsf{B}\mu \circ (\nu \circ \rho) : \phi$,
$\mathsf{B}(\mu + \nu) \circ \rho \leftrightarrow \mathsf{B}(\mu \circ \rho) + (\nu \circ \rho) : \phi$,
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.

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 into a structure of methods (built out of union and application) which is, so to speak, the scaffolding with which the neighbourhood functions for knowledge and belief are built. That is why methods models are more explanatory than neighbourhood ones, as we will see.

A neighbourhood model $\mathcal{F}$ is a pair $\langle W, N\rangle$ where W is a set of worlds and N a function from worlds to sets of propositions.[43]

Definition 15. 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.

A neighbourhood model is a pair $\langle \mathcal{F}, V\rangle$ where $\mathcal{F} = \langle W, N\rangle$ is a neighbourhood frame, with W a non-empty set of worlds and $N \subseteq W \to \mathcal{P}\mathcal{P}(W)$ a neighbourhood function, and $V : \mathsf{P} \to \mathcal{P}(W)$ a valuation function. We define $[\![\cdot]\!]^{\mathcal{M}}$:

$[\![\mathsf{p}]\!]^{\mathcal{M}} = V(\mathsf{p})$,
$[\![\neg\phi]\!]^{\mathcal{M}} = W \backslash [\![\neg\phi]\!]^{\mathcal{M}}$,
$[\![\phi \vee \psi]\!]^{\mathcal{M}} = [\![\phi]\!]^{\mathcal{M}} \cup [\![\psi]\!]^{\mathcal{M}}$,
$[\![\nabla\phi]\!]^{\mathcal{M}} = \{w : [\![\phi]\!]^{\mathcal{M}} \in N(w)\}$.
Truth. $\models^{\mathcal{M}}_w \phi$ iff $w \in [\![\phi]\!]^{\mathcal{M}}$.
Validity. $\models^{\mathcal{M}} \phi$ iff for any world w, $\models^{\mathcal{M}}_w \phi$.

In our models, methods are functions from worlds and sets of premises to sets of conclusions. For a given method m and set of premises π, 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. 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.

Theorem 21. *For each method frame, there is a pointwise equivalent neighbourhood frame for the simple language $\mathcal{L}_{\mathsf{B}}$, and conversely.*[44]

Proof. Let $\mathfrak{F} = \langle W, M^B, R\rangle$ be any methods frame. Define the neighbourhood frame $\mathcal{F} = \langle W, N\rangle$ such that for any w, $N(w) = B(w)$. We prove that $\mathcal{F}$ is pointwise equivalent to $\mathfrak{F}$ in $\mathcal{L}_{\mathsf{B}}$ 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 $\mathcal{F} = \langle W, N\rangle$ be any neighbourhood frame for B. Define the methods frame $\mathfrak{F} = \langle W, M^B, R\rangle$ 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 $M = \{m\}$. By the definition of union, $m + m = m$, and since m is non-inferential, $m \circ m = m$ (see Remark 4 in appendix A). Since $M^B = \{m\}$, $m + m = m$ and $m \circ m = m$, the application and union closure of M^B is $M = \{m\}$. We then prove that $B(w) = N(w)$. By the definition of m and by the definition of belief, $B(m, w) = m(w) = N(w)$ for any w. Since $M = \{m\}$, by the definition of belief again, $B(w) = N(w)$. From this $\mathcal{F}$ and $\mathfrak{F}$ are easily shown to be pointwise equivalent in $\mathcal{L}_{\mathsf{B}}$. □

Theorem 22. *Call a neighbourhood frame $\mathcal{F} = \langle W, N\rangle$ truthful iff for each w, $w \in \bigcap N(w)$, where $\bigcap N(w) = W$ whenever $N(w)$ is empty.*[45] *For each method frame, there is a pointwise equivalent truthful neighbourhood frame for the simple language $\mathcal{L}_{\mathsf{K}}$, and conversely.*

Proof. Let $\mathfrak{F} = \langle W, M^B, R\rangle$ be any methods frame. Define the neighbourhood frame $\mathcal{F} = \langle W, N\rangle$ such that for any w, $N(w) = K(w)$. We prove as before that the frames are pointwise equivalent in $\mathcal{L}_{\mathsf{K}}$. Moreover, we prove that $\mathcal{F}$ is truthful. For any w, if $N(w) = K(w)$ is empty, $\bigcap N(w) = W$ and $w \in \bigcap N(w)$; otherwise, we know that $w \in p$ for any $p \in K(w)$ (by the factivity of knowledge, Theorem 3), so $w \in \bigcap K(w)$ and $w \in \bigcap N(w)$.

Conversely, let $\mathcal{F} = \langle W, N\rangle$ be any truthful neighbourhood frame. Define the methods frame $\mathfrak{F} = \langle W, M^B, R\rangle$ 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 m is infallible at any

[44] $\mathcal{L}_{\mathsf{B}}$ is defined as $\mathcal{L}_{\nabla}$ above with the B operator replacing ∇. This corresponds to the fragment of $\mathcal{L}$ free of methods terms and the $\Box$ and K operators (Definition 1). Truth for $\mathcal{L}_{\mathsf{B}}$ in methods models is defined as in $\mathcal{L}$ without the idle clauses. $\mathcal{L}_{\mathsf{K}}$ (see below) is defined analogously.

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

world. Since $w \in \bigcap N(w)$, for any $p \in N(w)$, $w \in p$. Thus for any m, w, π, if $p \in m(w, \pi) = N(w)$, $w \in p$. Hence m is infallible at any world. Moreover $M = \{m\}$, so $K(w) = B(m, w) = N(w)$ for any w. From this we show as before that the frames are pointwise equivalent in $\mathcal{L}_{\mathsf{K}}$. □

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 21 and 22, 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 21 and 22, the class of methods frames is *not* equivalent to the class of neighbourhood frames for two modalities $\langle W, N^B, N^K \rangle$ where the N^K is truthful. For it is easy to see that in our models, $p \in K(w)$ entails $p \in B(w)$ for any w, p (Theorem 2), while we can construct neighbourhood models such that $p \in N^K(w)$ but $p \notin N^B(w)$ for some w, p. 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.

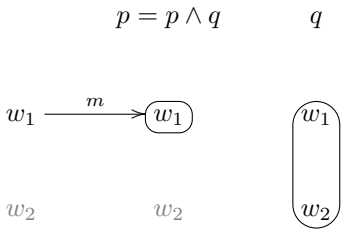
The same goes for other axioms. A much-discussed axiom for knowledge is **4**, according to which knowing is knowing that one knows: $Kp \rightarrow KKp$. A neighbourhood frame validates **4** iff: $p \in N^K(w) \rightarrow \{w' : p \in N^K(w')\} \in N^K(w)$ for any p, w. The condition is a transparent 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 the adequacy of an agent's methods is evaluated with a respect to an alethic modality such that what is possibly possible is possible, the agent satisfies axiom **4** for knowledge if she has "Confident Introspection" methods $pc(\{m\})$ such that, in non-formal terms: if she believes p out of m then she believes that she knows that p on a basis of $pc(\{m\})$ (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: one can easily find counterexamples to the axiom for agents who do not introspect or for a non-transitive space of

possibilities.

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 additional structure is reflected in the methods operators $\mathsf{B}\mu : \phi$ and $\mathsf{K}\mu : \phi$ and in theorems that cannot be stated with unary operators (Theorems 7, 8).

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 or know that q. Consider a frame $\mathfrak{F} = \langle W, M^B, R\rangle$ where $W = \{w_1, w_2\}$ with $p = \{w_1\}$, $q = W$ and $M^B = \{m\}$ where m is such that $m(w_1, \pi) = \{p\}$ and $m(w_2, \pi) = \emptyset$ for any π, and R any reflexive accessibility relation. 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_1.

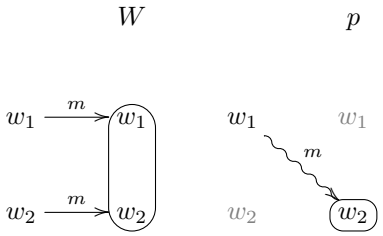


(The illustrations are explained in section 3.1.)

The methods frames $\mathfrak{F}$ constructed from a neighbourhood frame $\mathcal{F} = \langle W, N\rangle$ following the procedure in Theorem 22 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 13.) 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 there 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.

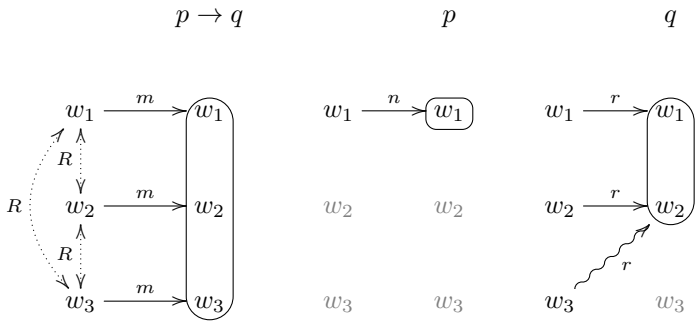
Example 2. Counterexample to $\mathbf{N_K}$. The agent has a method that leads her to believe both the tautology and a false proposition. Though the agent believes the tautology (W), she fails to know it, because her belief is fallibly based.

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 $w_1 \notin p$, $m \notin M^I(w_1)$. From this it follows that $\models^{\mathfrak{M}}_{w_1} \mathsf{B}\top$ but $\not\models^{\mathfrak{M}}_{w_1} \mathsf{K}\top$ (Definitions 6 and 7).



Example 3. Counterexample to $\mathbf{K}_{\mathsf{K}}$. We consider an agent that believes $p \to q$ and p by 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.



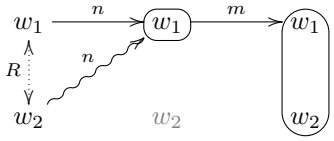
Consider $\mathfrak{M}$ in $\mathfrak{F}$ such that $V(\mathsf{p}) = p$ and $V(\mathsf{q}) = q$. We have $[\![\mathsf{p} \to \mathsf{q}]\!]^{\mathfrak{M}} = W$. Since $q \in r(w_3)$ but $w_3 \notin q$, and since $w_1 R w_3$, $r \notin M^I(w_1)$ (by Definition 4). It follows that at w_1 we have: $\models^{\mathfrak{M}}_{w_1} \mathsf{K}(\mathsf{p} \to \mathsf{q})$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Kp}$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Bq}$ 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 $p \to q$ 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 $p \to q$, and even if q was false.

Example 4. Counterexample to $\mathbf{4}_{\mathsf{K}}$. 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 $m \in M^I(w_1)$, $\models^{\mathfrak{M}}_{w_1} \mathsf{Kp}$. It is easy to check that $[\![\mathsf{Kp}]\!] = q$. Since $q = [\![\mathsf{Kp}]\!] \in n(w_1)$, $\models^{\mathfrak{M}}_{w_1} \mathsf{BKp}$. 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 6 and 7).

D Further exploration of Reasoning methods

Axioms M and C can be obtained separately. 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 16. *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 23. *($\boldsymbol{M}_{\mathsf{B}}$)* $\mathsf{B}(\phi \wedge \psi) \rightarrow (\mathsf{B}\phi \wedge \mathsf{B}\psi)$ *and ($\boldsymbol{M}_{\mathsf{K}}$)* $\mathsf{K}(\phi \wedge \psi) \rightarrow (\mathsf{K}\phi \wedge \mathsf{K}\psi)$ *are valid for any methods model such that $m^{SD} \in M$.*

($\boldsymbol{C}_{\mathsf{B}}$) $(\mathsf{B}\phi \wedge \mathsf{B}\psi) \rightarrow \mathsf{B}(\phi \wedge \psi)$ *and ($\boldsymbol{C}_{\mathsf{K}}$)* $(\mathsf{K}\phi \wedge \mathsf{K}\psi) \rightarrow \mathsf{K}(\phi \wedge \psi)$ *are valid for any methods model such that $m^{CD} \in M$.*

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

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 8 and 16. □

Correlation with the topology of sets. 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 $W \to \mathcal{PP}(W)$ 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}.

Axioms for belief and knowledge are independent. 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.

Reasoning methods are sufficient but not necessary for the logical omniscience axioms. 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, ϕ, ψ, $[\![\phi]\!] \in B(m, w)$ and $[\![\phi \to \psi]\!] \in B(n, w)$, there is some third method r such that $[\![\psi]\!] \in B(r, w)$, yet the third method is not the result of applying m^D to the union of m and n. 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 $p \to q$ from some persons, there happens to be, by sheer coincidence, a person that tells her that q. The agent thereby satisfies $\mathbf{K}_B$, $\mathsf{B}(\phi \to \psi) \to (\mathsf{B}\phi \to \mathsf{B}\psi)$. Yet that is intuitively accidental, because her belief that q is unconnected to her believing p and $p \to q$. 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 $p \to q$. 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 deduction preserves knowledge, and that generality can only be stated in the more complex language that allows reference to methods (see Theorem 8).

E Further exploration of Introspection methods

Model-relativity and constructing introspecter models. For any method model $\mathfrak{M}$ and set of methods X we have defined Introspection methods $pi(\mathfrak{M}, X)$ and $ni(\mathfrak{M}, X)$ and Confident Introspection methods $pc(\mathfrak{M}, X)$ and $nc(\mathfrak{M}, X)$. For a given methods model $\mathfrak{M}$, $pi(\mathfrak{M}, \cdot)$, $ni(\mathfrak{M}, \cdot)$, $pc(\mathfrak{M}, \cdot)$ and $nc(\mathfrak{M}, \cdot)$ can be seen as operations on sets of methods. But since they are only defined in model-relative terms (and in agent-relative terms in multi-agent settings), they are not model-independent operations such as union and application.

It is not straightforward to *construct* perfect introspection models, for related reasons. Let $\mathfrak{M}$ be a model where the agent lacks the corresponding Positive Introspection method: $pi(\mathfrak{M}, M) \notin M$. Suppose we want to build the "Introspection closure" of $\mathfrak{M}$. We cannot simply take the model $\mathfrak{N}$ with a method set N derived from the basic method set $M \cup \{pi(\mathfrak{M}, M)\}$. For it may be that $pi(\mathfrak{M}, M)$ is *not* an Introspection method at all in the new model: $pi(\mathfrak{M}, M) \neq pi(\mathfrak{N}, M)$. Even if we keep fixed the set of worlds between $\mathfrak{M}$ and $\mathfrak{N}$, it may be that (1) some method $m \in M$ outputs p at w, (2) the set of worlds where some method of the agent outputs p is not the same in $\mathfrak{M}$ than in $\mathfrak{N}$: $b(\mathfrak{M}, p) \neq b(\mathfrak{N}, p)$. If that is so, the output of $pi(\mathfrak{M}, M)$ that corresponds to the proposition that the agent believes that p in $\mathfrak{M}$ will not correspond to the proposition that the agent believes that p in $\mathfrak{N}$. The same holds for Negative Introspection. Thus there is no straightforward way to close a model under Introspection.

Multi-agent extension, common belief and common knowledge. Introspection methods are easily extended to multi-agent cases. Let $A = \{a, b, \ldots\}$ be a set of agents. Let M_a^B, M_a^B be their basic methods sets and M_a, M_b be their methods set, built as in the single-agent case (Definition 5). The functions for belief and knowledge are parametrised accordingly: $B(a, m, w)$, $B(a, w)$, $K(a, m, w)$, $K(a, w)$. Relative to a model $\mathfrak{M}$, the proposition that agent a believes p is: $b(\mathfrak{M}, a, p) = \{w | p \in B(a, w)\}$ (see Definition 10). Agent-relative Positive Introspection is defined thus (see Definition 11):

Definition 17. Given a model $\mathfrak{M}$, an agent a and a set of methods X, the *Positive Introspection of X for a in $\mathfrak{N}$*, notated $pi(\mathfrak{N}, a, X)$, is the method such that for any w, p, π: $p \in pi(\mathfrak{M}, a, X)(w, \pi)$ iff for some p', $p = b(\mathfrak{M}, a, p')$ and there is a $m \in X$ such that $p' \in m(w)$.

Whenever some m in X outputs p, $pi(\mathfrak{M}, a, X)$ outputs that a believes p. In rough terms, an agent that has $pi(\mathfrak{M}, a, X)$ takes a to have at least the methods in X, as far as a's beliefs are concerned. Thus these "Introspection" methods become "Mind-Reading" methods when they belong to another agent than the introspected one. Similar extensions of ni, pc and nc can be made. We can establish analogues of Lemmas 1 and 2 and, with transitivity, Lemma 4. For instance:

Lemma 5. *For any multi-agent methods model, set of methods X and agent a, if $X \subseteq M_a$, then the Positive Introspection of X for a in that model is infallible.*

Notions of common belief and common knowledge can be derived. Say that a method m is *shared* in a group G iff $m \in M_a$ for each agent $a \in G$:

Definition 18. A set of methods X is *shared and commonly introspected in G* iff $X \subseteq M_a$ for every $a \in G$ and $pi(a, X) \in X$ for every a in G.

A set of methods is *shared and commonly confidently introspected in G* iff $X \subseteq M_a$ for every $a \in G$ and $pc(a, X) \in X$ for every a in G.[46]

It is easy to see that if X is shared and commonly introspected in G, whenever some agent in a in G believes p on the basis of some method in X, all the agents in G believe that a believes p. Moreover, they do so on the basis of an infallible method (Lemma 5), so they all know that a believes p. Thus the beliefs of agents in G based on methods in X are common. Moreover, if all the methods in X are infallible, and if we assume transitivity, $pc(a, X)$ is infallible for any $a \in G$ (see Lemma 4). It follows that if X is shared and commonly confidently introspected in G, whenever some agent a in G knows something on

[46] $pc(\mathfrak{M}, a, X)$ is the multi-agent extension of Confident Introspection that parallels the multi-agent extension of Introspection above.

the basis of a method in X, all the agents in G know that a knows it. Thus the knowledge of agents in G based on methods in X is common.

The grain of introspection methods. To derive knowledge of one's knowledge, we relied on very fine-grained Positive Confident Introspection methods: one method $pc(\{m\})$ for each method m of the agent. An indiscriminate confidence method $pc(M)$ would be fallible as long as some of the agent's methods are fallible, which would block a proof of knowledge of one's knowledge. These results are intuitive. Believing that one knows on the basis of $pc(M)$ means believing it on the mere basis that one believes. If some of one's beliefs are false, this is not good enough to know that one knows. By contrast, believing that one knows on the basis of the finer-grained method $pc(m)$ is tantamount to believing it *because one believes on the basis of* m. If m is infallible, this is a better ground.[47]

We could have derived knowledge of one's knowledge with less fine-grained methods: it is sufficient that the agent has each $pc(X)$ where X is the set of methods that are infallible at some world w ($X = M \cap M^I(w)$ for some w). It follows that at each world, the agent has a confidence method that introspects exactly those methods that are infallible at that world. For that knowledge of one's knowledge is easily derived, assuming transitivity. But it seems to me arbitrary that the grain of an agent's higher-order methods should match the way infallibility partitions them at various worlds. The idealisation under which an agent's higher-order methods are simply just as fine-grained as the lower-order ones is more natural.

By contrast, we could not have derived negative introspection and negative confident introspection results with finer-grained methods $ni(\{m\})$ and $nc(\{m\})$. Such methods tell that the agent fails to believe (or know) p whenever m fails to output p. There are typically fallible if the agent has other methods than m. As we pointed out (section 5.4.1), the asymmetry is due to the fact that belief merely requires the output of *some* method of the agent, while absence of belief requires the absence of an output from *all* methods of the agent. However, we may have opted for a more holistic notion of belief, under which an agent believes that p provided that one of her methods output p *and no other method outputs* $\neg p$, for instance (section 4.3.3). On such a notion Positive Introspection would have to be defined holistically as well.

Confidence should be inferential. Our psychological Introspection methods pi and ni are perceptual-like: they are non-inferential. The simple fact that

[47] Not a sufficient one, yet: it need also be the case that believing that one knows because one believes on the basis of m is itself an infallible method, which requires in turn m to be also infallible at accessible worlds — hence the transitivity requirement.

a method m in X outputs p at a world makes the method $pi(X)$ unconditionally output the proposition that the agent believes p. Thus the methods $pi(X)$ and $ni(X)$ "track" what the methods in X are doing in the manner in which perceptual methods track external facts. When the methods in X are those of the agent, the Introspection of X quite literally perceives the inner.

Our Confident Introspection methods also track methods outputs in a perceptual-like manner. That is why there are Introspection methods as well. But it would be more natural to get them as the result of applying an *inferential* Confidence method to the outputs of psychological Introspection methods. The idea is this: whenever the agent believes p, she believes that she believes p (psychological introspection) and whenever she believes that she believes p, she infers that she knows it (inferential confidence).

That can be done for Negative Confidence: we can define for each model a method nc^* such that $nc^*(w, \pi) = \{-k(p) : -b(p) \in \pi\}$. It is easy to show that the method is infallible. Since $ni(M)$ is infallible, the composed method $nc^* \circ ni(M)$ is infallible and we can derive (**p5**) for Perfect Introspecters with nc^*. In addition, we get the schema $\mathsf{B}\neg\mathsf{B}\phi \rightarrow \mathsf{B}\neg\mathsf{K}\phi$ 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. That is a consequence of the individuation of propositions as sets of possible worlds. To illustrate, suppose we have fine-grained psychological introspection: $pi(\{m\})$ for each $m \in M$. Let m be some infallible method of the agent. For each p that m outputs at w, $pi(\{m\})$ outputs $b(p)$ at w. Now suppose we simply define the positive confidence method pc^* such that $pc^*(w, \pi) = \{k(p) : b(p) \in \pi\}$. Then $pc^* \circ pi(\{m\})$ will output each $k(p)$ where p is an output of m — and since m is among the agent's infallible methods, all these outputs of $pc^* \circ pi(\{m\})$ will be true, as desired. But the problem is that $pc^* \circ pi(\{m\})$ may output *more* than that. For it may be that the worlds at which m outputs p are precisely those at which some other method n outputs q, and that it turns out that $b(p) = b(q)$. Since $b(p) \in pi(\{m\})(w)$, $b(q) \in pi(\{m\})(w)$, and by the definition of pc^*, pc^* will *also* output $k(q)$. But if n is the only method that outputs q and if it is fallible, $k(q)$ will be false. So $pc^* \circ pi(\{m\})$ may be fallible even if m is infallible, which blocks the derivation of knowledge of one's knowledge. Intuitively, this is so because the inferential method could not differentiate between the proposition that the agent believes that p and the proposition that she believes that q. Various refinements of Positive Confidence could be envisaged to circumvent the problem, but we leave this issue to further work.[48]

[48] Does the problem not affect inferential Negative Confidence as well? Yes, but there it is not an obstacle to the infallibility of $nc^* \circ ni(M)$. Suppose $ni(M)$ outputs $-b(p)$, but that it

The role of transitivity and euclideanity. It is *prima facie* surprising that knowledge of one's knowledge requires an assumption on the background modality, while knowledge of one's ignorance (partial or not) does not.[49] That is cleared up when we reflect of the fact that there is only one way to know, but *two* ways to fail to know. To know p, one must have a method that outputs p and that is infallible. Since infallibility depends on what happens at accessible worlds, knowledge does as well. To fail to know p, one must *either* fail to believe it, or believe it only on the basis of fallible methods. Fallibility depends on what happens at accessible worlds, so the second way of failing to know does as well. But the first way does not. Knowledge is always a modal matter, but ignorance is a modal matter only when it consists in fallible belief.

As a result, if a method outputs at some accessible world distinct from the actual one that the agent knows that p, whether that method is infallible at the actual world will depend on what goes on at two steps of accessibility. (If it outputs that the agent knows that she knows, that will depend on what goes on at three steps, and so on.) By contrast, if a method outputs at some non-actual accessible world that the agent *fails* to know that p, its infallibility will depend on what goes on at two steps of accessibility *only if some such failure results from fallible belief*. But our Negative Confident introspection methods only output that the agent fails to know that p when they fail to believe it. Hence their infallibility only depends on what goes on at one step of accessibility. That is why we get (partial) knowledge of one's ignorance without further assumption on accessibility.

We could imagine a *Modesty* method that mirrors positive Confident Introspection, such that when a certain methods outputs something, the agent believes that they do *not* know it. Modesty about a method amounts to taking that method to be fallible. If the original method is actually fallible, Modesty about it only yields true outputs at the actual world. But for Modesty to be *infallible*, the original method should also be fallible at all accessible worlds. Now if the background accessibility relation is euclidean, a method that is fallible at a world must be fallible at worlds accessible from it (compare with Lemma 3).[50] Knowledge of one's ignorance through Modesty would thus require a constraint on accessibility parallel to the one familiar from epistemic logic, in which axiom

turns out that $-b(p) = -b(q)$. Then $nc^* \circ ni(M)$ will output $-k(p)$ but also $-k(q)$. However, since $ni(M)$ is infallible, $-b(p)$ is true; and since $-b(p) = -b(q)$, $-b(q)$ is true as well. So there is no risk that $-k(q)$ is false. By contrast, in the analogous Positive case, the truth of $b(q)$ does not guarantee the truth of $k(q)$, so the inference may fail.

[49] Thanks to Martin Smith for raising the issue.

[50] R is euclidean iff whenever a world is accessible, it is also accessible from all the accessible worlds: $\forall w, w', w''((wRw' \wedge wRw'') \rightarrow w''Rw')$. Suppose m is fallible at w: then there is a world w' where it outputs a false proposition, conditional on some set of true premises. Suppose w'' is accessible from w: if R is euclidean, w'' has access to w' as well, and since m has a false output in w', m is fallible in w'' as well.

5 requires euclideanity.[51]

F Belief, knowledge and information

Information. As understood here, information is an objective notion of content. The information contained in a proposition is just the set of possibilities compatible with it: namely, the set of worlds in which it is true. Similarly, the information contained in a *set* of propositions is the set of worlds that are compatible with each proposition in the set. As a special case, the empty set of propositions is compatible with any world: its information is the same as that of a tautology. A set of propositions containing a contradiction or contradictory propositions is not compatible with any world, so its information is the empty set. With this notion we can define the (conditional or unconditional) information provided by a method at a world as the set of possibilities compatible with the (conditional or unconditional) outputs of that method at that world, and the agent's doxastic and epistemic information as the set of possibilities compatible with what an agent believes and with what she knows, respectively.

Definition 19. For any set of propositions π, we define:

$I(\pi) = \{w \in W | \forall p(p \in \pi \rightarrow w \in p)\}$.

For any method m, world w and set of premises π:

$I(m(w, \pi))$ is the *information provided by* m *at* w *on the basis of* π.

For any methods model $\mathfrak{M}$ and world w:

$I(B(w))$ is the agent's *doxastic information* at w, and

$I(K(w))$ is the agent's *epistemic information* at w.

Note that the information of a method is defined independently of whether the agent has it; by contrast, doxastic and epistemic information are agent- and model-dependent.

The relations between $I(m(w, \pi))$ and $I(B(w))$ and $I(K(w))$ are straightforward. The agent's doxastic information is the unconditional information given by all the methods she has: $I(B(w)) = \bigcap_{m \in M} I(m(w))$, provided that M is not empty. (If M is empty, $I(B(w)) = W$.) The agent's epistemic information is the unconditional information given by the infallible methods she has: $I(K(w)) = \bigcap_{m \in M \cap M^I(w)} I(m(w))$, provided $M \cap M^I(w)$ is not empty, otherwise $I(K(w)) = W$.

[51] Note that euclideanity would not be sufficient to make Modesty infallible conditional on the fallibility of the original methods. Modesty about at set of methods X outputs that the agent does not know that p whenever some $m \in X$ outputs p. This will be the case only if m is fallible *and no other infallible method of the agent outputs* p. An adequate Modesty about X should only output that the agent does not know that p when no other method of the agent that those in X output that p. This would make Modesty holistic in the way Negative Introspection and Confidence are.

Explicit and Implicit belief and knowledge. The formal representation of such notions is familiar from Hintikka (1962) models. The worlds compatible with what I believe are just those worlds where every proposition I believe holds, and similarly for knowledge. Our notions of information thus correspond to standard Kripke models:

Definition 20. *Information models.* For any frame $\mathfrak{F}$, let $R^B, R^K \subseteq W \times W$ be such that, for any w, w':

wR^Bw' iff $w' \in I(B(w))$,

wR^Kw' iff $w' \in I(K(w))$.

Each of $\langle W, R^B\rangle$, $\langle W, R^K\rangle$ is a Kripke frame. We can introduce corresponding modal operators D and E in the language. For any methods model $\mathfrak{M}$,

$[\![\mathsf{D}\phi]\!]^{\mathfrak{M}}=\{w|\forall w'(wR^Bw' \to w \in [\![\phi]\!]^{\mathfrak{M}})\}$,

$[\![\mathsf{E}\phi]\!]^{\mathfrak{M}}=\{w|\forall w'(wR^Kw' \to w \in [\![\phi]\!]^{\mathfrak{M}})\}$.

Corollary 5. $\mathsf{B}\phi \to \mathsf{D}\phi$ *and* $\mathsf{K}\phi \to \mathsf{E}\phi$ *for any methods model.*

Proof. Let $\mathfrak{M}, w$ be such that $\models^{\mathfrak{M}}_w \mathsf{B}\phi$. By the definition of belief, $[\![\phi]\!] \in B(w)$. By the definition of information, $I(B(w)) \subseteq [\![\phi]\!]$, and by Definition 20, $\forall w'(wR^Bw' \to w \in [\![\phi]\!])$, so $\models^{\mathfrak{M}}_w \mathsf{D}\phi$. And similarly for $\mathsf{K}\phi \to \mathsf{E}\phi$. □

By contrast, it easy to find models in which $\mathsf{D}\phi \to \mathsf{B}\phi$ fails: for instance, $\mathsf{D}\top$ holds at any model, but $\mathsf{B}\top$ fails at some. This gives a sense in which D and E represent the information contained in one's belief and knowledge: explicitly, when $\mathsf{D}\phi \wedge \mathsf{B}\phi$, and implicitly, when $\mathsf{D}\phi \wedge \neg\mathsf{B}\phi$, and similarly for E and K.

Informativeness. Intuitively, a method is informative iff it can reduce the set of possibilities the agent considers or should consider. Formally, we define:

Definition 21. A method m is *uninformative* iff $I(m(w, \pi)) \supseteq I(\pi)$ for all w, π. A method is informative iff it is not uninformative.

In particular, m is uninformative only if its unconditional outputs are not more informative than the tautology: $I(m(w)) \supseteq I(\emptyset) = W$ for any w. A method that has some unconditional output other than the tautology is informative.

The union of a uninformative method m with a method n does not give more information than is contained in premises or given by n. The application of an uninformative method m to n does not give more information than n did not already give. Formally:

Corollary 6. *If m is an uninformative method, then for any n, w, π:*

$I(m + n(w, \pi)) \supseteq I(\pi \cup n(w, \pi))$,

$I(m \circ n(w, \pi)) \supseteq I(n(w, \pi))$.

Proof. Note first that $I(\pi \cup \pi') = I(\pi) \cap I(\pi')$ for any π, π'. For by the definition of information, $w' \in I(\pi \cup \pi')$ iff $\forall p((p \in \pi \vee p \in \pi') \rightarrow w' \in p)$, or equivalently, $\forall p(p \in \pi \rightarrow w' \in p) \wedge \forall p(p \in \pi' \rightarrow w' \in p)$, that is $w' \in I(\pi) \cap I(\pi')$. From this and the definition of union, we have $I(m + n(w, \pi)) = I(m(w, \pi)) \cap I(n(w, \pi))$ for any m, n, w, p. Now if m is uninformative, $I(m(w, \pi)) \supseteq I(\pi)$ (Definition 21), so $I(m(w, \pi)) \cap I(n(w, \pi)) \supseteq I(\pi) \cap I(n(w, \pi))$.

By the definition of application, $m \circ n(w, \pi) = m(w, n(w, \pi))$ for any m, n, w, π. And if m is uninformative, $I(m(w, n(w, \pi))) \supseteq I(n(w, \pi))$ (Definition 21). □

Union and application of uninformative methods are uninformative:

Corollary 7. *If m and n are uninformative methods, $m + n$ and $m \circ n$ are uninformative.*

Proof. Suppose m and n are uninformative. By Corollary 6, $I(m + n(w, \pi)) \supseteq I(\pi \cup \pi) = I(\pi)$ for any w, π. By Corollary 6 again, $I(m \circ n(w, \pi)) \supseteq I(n(w, \pi))$ for any w, π, and since n is uninformative, $I(n(w, \pi)) \supseteq I(\pi)$. □

Uninformative methods are guaranteed to preserve truth, in the following sense:

Corollary 8. *Say that a method m is truthful at w iff $w \in I(m(w))$. If a method m is uninformative, then for any n, $m + n$ and $m \circ n$ are truthful at w if n is.*

Proof. Suppose n is truthful at w: $w \in I(n(w))$. If m is uninformative, $I(m + n(w)) \supseteq I(n(w))$ (Corollary 6), so $w \in I(m + n(w))$. Again, if m is uninformative, $I(m \circ n(w)) \supseteq I(n(w))$ (Corollary 6), so $w \in I(m \circ n(w))$. □

This characterises the sense in which uninformative methods are risk-free. Uniting an uninformative method with another or applying it to another cannot lead to false beliefs unless the original method did.

Theorem 24. *Deduction and Pure-Reason are uninformative.*

Proof. By Definition 8, $m^R(w, \pi) = \{W\}$ for every w, π. So $I(m^R(w, \pi)) = W \supseteq I(\pi)$ for any w, π.

By Definition 8, $m^D(w, \pi) = \{p | \exists q, r \in \pi(p \supseteq q \cap r)\}$. Suppose $w' \in I(\pi)$ for some π; then $w' \in q \cap r$ for any $q, r \in \pi$ (by the definition of information), so $w' \in I(m^D(w, \pi))$ at any w. □

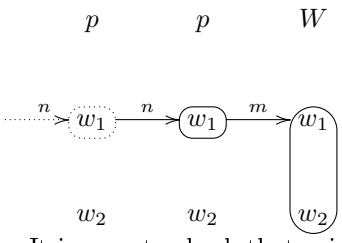
Theorem 25. *Say that a method m is included in a method n iff for all w, π, $m(w, \pi) \subseteq n(w, \pi)$. The most inclusive uninformative method is the Total Reasoning method m^{TR} such that $m^{TR}(w, \pi) = \{p | p \supseteq I(\pi)\}$ for any w, π.*

Proof. Let m be a method not included in m^{TR}. Then there is some p, w, π such that $p \in m(w,\pi)$ and $p \not\supseteq I(\pi)$. So for some w', $w' \in I(\pi)$ but $w' \notin p$. Hence $w \notin I(m(w,\pi)$ and $I(m(w,\pi)) \not\supseteq I(\pi)$, so m is informative. □

By contrast, Introspection methods (section 5.4) are typically informative.

Theorem 26. *Introspection and Confident Introspection are informative. For some methods models, there are w, π such that $I(pi(M)(w,\pi)) \not\supseteq I(\pi)$, and similarly for $ni(M)$, $pc(\{m\})$ for $m \in M$, $nc(M)$.*

Proof. Model for $I(pi(M)(w,\pi)) \not\supseteq I(\pi)$. Let $W = \{w_1, w_2\}$, $p = \{w_1\}$, $q = \{w_2\}$ and $M = \{m, n\}$ where m is such that for any π, $m(w,\pi) = \{W\}$ if $w = w_1$ and $\emptyset$ otherwise, and n such that for any π, $n(w,\pi) = \{p\}$ if $w = w_1$ and $\emptyset$ otherwise.



It is easy to check that n is the Positive Introspection of M, $pi(M)$. At w_1, m outputs W and n outputs p. Correspondingly, n outputs the set of worlds in which some method in M outputs W, namely $\{w_1\}$, and the set of worlds in which some method in M outputs p, namely $\{w_1\}$ as well. (Since $\{w_1\} = p$, n introspects itself.) At w_1 we have $I(n(w_1)) = p \not\supseteq I(\emptyset) = W$. Similar models can be built for $ni(M)$, $pc(\{m\})$ for $m \in M$, $nc(M)$. □

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 that p typically narrows down the sets of worlds compatible with the agent's belief. Similarly, p and Kp do not typically hold at the same worlds. That is why Introspection and Confident Introspection are informative methods. 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, London, 1973.

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

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

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

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

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.

John Hawthorne and Maria Lasonen-Aarnio. Knowledge and objective chance. In Peter Greenough and Duncan Pritchard, editors, *Williamson on Knowledge*, pages 92–108. Oxford University Press, 2009.

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.

Angelika Kratzer. The notional category of modality. In H.-J. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Contexts*, pages 38–74. de Gruyter, Berlin, 1981.

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

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

Marua Lasonen-Aarnio. Single premise deduction and risk. *Philosophical Studies*, 141(2):157–173, 2008.

David Lewis. Attitudes *De Dicto* and *De Se*. In *Philosophical Papers*, volume 1, pages 133–155. Oxford University Press, 1979/1983.

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.

Willard van Orman Quine. Propositional objects. In *Ontological Relativity and Other Essays*, pages 139–167. Columbia University Press, 1969.

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.

Martin Smith. What else justification could be. *Noûs*, 44:10–31, 2010.

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.

J. Robert G. Williams. Chances, counterfactuals, and similarity. *Philosophy and Phenomenological Research*, 77(2):385–420, 2008. doi: 10.1111/j.1933-1592.2008.00196.x. URL http://dx.doi.org/10.1111/j.1933-1592.2008.00196.x.

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

Timothy Williamson. Probability and danger. *The Amherst Lecture in Philosophy*, pages 1–35, 2009. URL http://www.amherstlecture.org/williamson2009/.