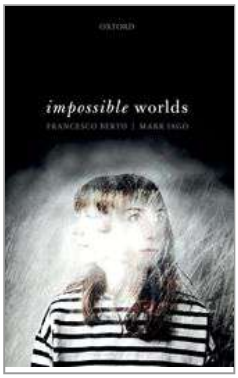



Impossible Worlds
Francesco Berto and Mark Jago

Print publication date: 2019
Print ISBN-13: 9780198812791
Published to Oxford Scholarship Online: August 2019
DOI: 10.1093/oso/9780198812791.001.0001

# Epistemic and Doxastic Contents

Francesco Berto
Mark Jago

DOI:10.1093/oso/9780198812791.003.0010

## Abstract and Keywords

The case for making belief states the primary focus of our analysis and for including impossible worlds in that analysis is outlined in this chapter. This allows the reader to deny various closure principles, although this won't help defeat worries about external-world scepticism. The issue that concerns the authors most is the *problem of bounded rationality*: belief states seem to be closed under 'easy' trivial consequence, but not under full logical consequence, and yet the former implies the latter. The solution presented here is that some trivial closure principle must fail on a given belief state, yet it is indeterminate just where this occurs. Formal models of belief states along these lines are given and it is shown that they respect the indeterminacy-of-closure intuition. Finally, the chapter discusses how we might square this approach with the fact that some people seem to believe contradictions.

*Keywords:* closure principles, bounded rationality, trivial consequence, indeterminacy, contradiction, scepticism

## 10.1 Belief States

Analyses of belief often focus primarily on individual beliefs, analysed as a relation between an agent and a sentence or proposition. But this isn't the only approach. One may, instead, focus on the agent's total belief state. That approach is appealing if we follow Dennett (1987) and Stalnaker (1984) in analysing belief as a dispositional, functional state, essentially tied to action. Recall (§8.2) Stalnaker's idea that

> Representational mental states should be understood primarily in terms of the role that they play in the characterization and explanation of action. … Our conceptions of belief are conceptions of states which explain why a rational agent does what he does.
>
> (Stalnaker 1984, 4)

An agent who is disposed to act in certain ways in a range of possible circumstances thereby has certain beliefs and desires. Being so disposed to act is precisely what it is to have those beliefs and desires, on this view. But we can't assign belief-desire pairs on the basis of single actions. There's any number of belief-desire explanations of Anna's taking an umbrella this morning. (Does she want to avoid getting wet? Does she want a makeshift weapon to hand? Does she merely like the look of carrying an umbrella, whatever the weather?)

When we factor in the agent's overall pattern of behaviour, by contrast, we can attempt a general rational explanation by ascribing **(p.214)** beliefs and desires. The important point here, for present purposes, is that belief is ascribed holistically, as part of a rational explanation of the agent's behaviour. So a philosophical explanation of belief should begin with holistic belief states. Individual beliefs are found further downstream.

The worlds approach model of belief (which we met in §1.2 and Chapter 5) naturally gives us an analysis of a total belief state, in the first instance. An agent's total belief state can be conceptualized, at a high level of abstraction, as a function on circumstances. The function categorizes those circumstances, depending on whether the agent can rule out a given circumstance. A state of complete ignorance corresponds to a function which, like a huge shrug of the shoulders, rules no circumstance out. Acquiring some information (or at least, what the agent takes to be information) corresponds to ruling out some of these potential circumstances. The agent's state is effectively saying, things aren't like *that*. (This way of thinking about things naturally ties in with the worlds approach to information from Chapter 9. We'll say more about the relationship in §10.5.)

This is a very general approach to thinking about representational states, and we can use it to conceptualize both knowledge and belief. Let's revisit our weather example from §9.2, in which we discussed the Bar-Hillel-Carnap theory of semantic information (Bar-Hillel and Carnap 1953). Suppose you consult the UK Met Office website and discover that it often rains in Manchester, something you previously didn't know. You then form the belief that it often rains there and, let's suppose, this belief counts as knowledge. Prior to consulting the Met Office, you had no idea about Manchester's average rainfall. As far as you were concerned, it might rain seldom there, as in Cambridge. But now you know better: Manchester is much wetter than Cambridge.

In your previous state, scenarios in which Manchester's climate is drier than Cambridge's were *epistemic possibilities* for you. If asked whether Manchester is as those scenarios say it is, you might have said 'it might be, for all I know'. This shrugging, non-committal attitude towards a scenario is what we mean by saying that the scenario is epistemically possible for you at that time. The epistemically possible **(p.215)** scenarios are the ones that represent a way the world might well be, for all you know.

Now that you know that it often rains in Manchester, however, those scenarios in which Manchester is mostly dry are no longer epistemically possible for you. In gaining your new piece of knowledge, you ruled out those scenarios as ways the world might be. Gaining new knowledge goes hand in hand with ruling out various scenarios as ways the world might be, as far as you're concerned.

Bar-Hillel and Carnap (1953), Chalmers (2002a, 2010), Hintikka (1962), Lewis (1975, 1986b), and Stalnaker (1976b, 1984) all develop a variation on this basic idea. It forms the basis of the Bar-Hillel-Carnap theory of semantic information (1953), on which information is conceptualized in terms of excluding certain worlds. Hintikka's (1962) semantics for epistemic logic (§5.1) can be seen in a similar light. Its crucial innovation was to add *epistemic accessibility* relations to the space of worlds. These relativize the process of ruling out worlds to specific worlds: the worlds are ruled out, from the perspective of world $w$, may differ from the worlds ruled out from the perspective of some other world $w_1$. As we put things in §5.1: the worlds accessible from $w$ need not be those accessible from $w_1$.

We analyse an agent's total state of knowledge or belief, at world $w$, in terms of all worlds accessible from $w$. We can recapture individual beliefs or bits of knowledge by looking at what those worlds represent in common. Agent $i$ believes that $A$ (at $w$) when all worlds accessible from $w$, for $i$, represent that $A$. One very happy aspect of the worlds approach is the way the analyses of information and of belief (and knowledge) interact. In Chapter 9, we conceptualized information in terms of a partition on worlds, and we've just conceptualized belief in terms of epistemic accessibility relations on worlds.

How do these two approaches interact? Gaining information leads to new beliefs (and, in the right circumstances, to new knowledge). We can conceptualize this dynamic process in terms of an *update* on the agent's epistemic accessibility relations. The content of a learned piece of information restricts the agent's accessibility relations to the worlds contained in that content. The dynamics of *information update* and **(p.216)** the subsequent effect on belief, which we briefly introduced in §5.5, was studied by Baltag et al. (1998), Segerberg (1995), Van Benthem (2011), Van Ditmarsch (2005), and Van Ditmarsch et al. (2008). (Baltag and Renne (2016) give an overview of the field.)

What we want to bring out here is the deep connection between the philosophical view of belief as a dispositional, functional state, and the worlds semantics. In making the case for the semantics, we may appeal to its usefulness in the sciences as well as to its philosophical underpinning. Not many theories have it so good, and we shouldn't pass up on those benefits lightly.

## 10.2 The Impossible Worlds Approach

The main drawback when this approach to belief is presented in terms of possible worlds is the *logical omniscience problem*, which we discussed in some detail in §5.1. The objection hits the philosophical analysis of belief states just as hard. By now, our response should be no surprise: we suggest an analysis of belief (and knowledge) states in terms of possible and impossible worlds. That is, we maintain the analysis of belief (and knowledge) in terms of epistemically possible scenarios, but we do not require that these be metaphysically or logically possible. What seems possible, from the cognitively limited perspective of a real-world agent, need not be metaphysically or even logically possible. Epistemically accessible worlds must *seem* possible to the agent in question but, as we discussed in §9.5, an impossible world may well seem possible.

On reflection, there's nothing in the philosophical motivation for the dispositional, functional account of belief states which rules out this impossible worlds approach. On Stalnaker's view, we analyse belief states ultimately in terms of our conception of rational action, where:

> What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions.
>
> (Stalnaker 1984, 4)

**(p.217)** Although 'rational' plays an essential role in Stalnaker's explanation, 'possible' does not. All the work is done by 'alternative outcomes', so long as we understand the range of alternatives in question in a broad enough way. (We'll discuss whether the impossible worlds approach adequately captures the *rational* component in §10.3.)

We're going to discuss two philosophical features of the impossible worlds approach to belief states: what it says about Frege's problem, which we presented in §9.3, and what it says (or doesn't say) about scepticism.

In §9.3, we asked how it is that 'James Newell Osterberg is Iggy Pop' is informative, when 'Iggy Pop is Iggy Pop' is not. There is an analogous problem for belief and knowledge: how can one believe (or know) that Iggy Pop has turned 70, without thereby believing (or knowing) that Osterberg has turned 70?

Since Osterberg is Iggy Pop, why are we not entitled to replace ‘Iggy Pop’ with ‘Osterberg’, inferring belief (or knowledge) that Osterberg has turned 70?

In general, Frege’s problem of belief concerns the failure of the inference from ‘$x$ believes that $Fa$’ and ‘$a = b$’ to ‘$x$ believes that $Fb$’. The problem is to account for the failure, within a semantically plausible general theory.

The impossible worlds approach solves the problem smoothly. Suppose that $a$ is $b$. Then it’s necessary that $a$ is $b$, and so metaphysically impossible that $a$ and $b$ are distinct. Given Nolan’s Principle (or one of the stronger principles) from §8.4, we infer that there are impossible worlds according to which $a$ isn’t $b$. We may then reasonably assume that there is an impossible world $w$ according to which $a$ is $F$, but which doesn’t say that $b$ is $F$. (This is implied directly if we assume either (8.6) or ($NP^+$) from §8.4.) Then, by taking world $w$ to be an epistemic possibility, we can model agents as believing that $Fa$ but not that $Fb$. On this approach, the inference from ‘$x$ believes that $Fa$’ and ‘$a = b$’ to ‘$x$ believes that $Fb$’ fails.

This solution falls out of the impossible worlds approach in a natural way. If an agent believes something under her $a$-concept, but not under her $b$-concept, then it must be acceptable both to theorize in terms of $a$-representations which are not $b$-representations, as in **(p.218)** the case of our world $w$, and to take that world to be epistemically accessible for that agent. For in general, we informally understand ‘epistemically accessible world’ in terms of what could be the case, for all the agent knows. If she doesn’t know that $a$ is $b$, then some world according to which $a$ isn’t $b$ is accessible to her. The solution falls out of our general understanding of epistemic accessibility.

Now let’s consider what consequences, if any, the impossible worlds approach (as applied to knowledge) has for scepticism. You know that, if you’re currently reading this book, then you’re not in some devilish sceptical scenario. You’re not having some super-realistic dream whilst really in bed, you haven’t been reduced to a brain hooked up to a hallucination-machine by an evil scientist, or anything like that. These are good things to know. But the antecedent is correspondingly hard to know. For if you know that, then you know that you’re safe from scepticism.

The key premise in this kind of reasoning, which went by without much fanfare, is *Closure under Known Implication*, one version of which we already met in §5.1 as closure condition (C6):

> (C6) If an agent knows that $A$ and knows that $A$ implies $B$, then she knows that $B$.

Those of a knowledge-positive disposition then reason by *modus ponens*. You know you're reading this book. So, you infer, you're not in a sceptical scenario. Scepticism is defeated! But this seems too easy. How, given the mere possibility of sceptical scenarios, are you so sure that you know you're in fact reading this book? The knowledge-negative alternative is to reason by *modus tollens*: since you don't know that you're not in a sceptical scenario, you don't know that you're currently reading this book. These alternatives, *dogmatism* via *modus ponens* on (C6) or *scepticism* via *modus tollens*, seem equally unattractive.

A third approach is to deny (C6). Then you seem to have the best of both worlds: knowing that you're reading this book, even though you can't in general rule out sceptical scenarios. Both Dretske (1970) **(p.219)** and Nozick (1981) make this move. According to Dretske, Mark knows he's been writing with Franz, even though he can't completely rule out a (wildly implausible) alternative, featuring a convincing Franz-impersonator, set on co-authoring a book about impossible worlds. According to Nozick, Franz knows where his house keys are, even though he can't completely rule out the (wildly implausible) scenario in which they've been stolen and replaced by useless replicas.

For Nozick, one's belief (and for Dretske, one's evidence) that $A$ must be sensitive in the proper way to the truth of $A$. They express this idea counterfactually: had $A$ been false, one would not have believed (have had evidence for) $A$. Had Mark not been writing with Franz, he wouldn't have believed he was. (That's true because the counterfactual selects the closest antecedent worlds, and these are worlds free of Franz-impersonators, in which Mark and Franz never agree to write together. Franz-impersonators belong to distant worlds.) In each case, (C6) fails. Good riddance, say Dretske and Nozick.

An impossible worlds approach to knowledge can be a hostile environment for closure principles on knowledge, including (C6). We showed in Chapter 5 how impossible worlds can be put to use to invalidate closure principles. All that's needed for a closure principle to fail is a single epistemically possible world in which the known premises hold but the putatively known conclusion does not. In particular, all that's needed for (C6) to fail is a world not closed under *modus ponens* to count as epistemically possible for some possible agent. We argued in §8.2 that there are such epistemic possibilities, and so we deny (C6). The stage seems set for us to add 'solving scepticism' to our list of benefits of the impossible worlds approach.

That would be far too quick. Epistemic closure fails, we've argued, because epistemic agents lack the cognitive resources to draw out all consequences of what they know. Take our chess players from §8.2. One player has a winning strategy available to her, yet doesn't know how best to proceed. That's because working out the winning strategy is just too complex. Compare this with a task the agent can *easily* compute. Nothing we've said so far speaks to *that* case. We

**(p.220)** deny closure in hard cases, but this leaves open the possibility of closure for easy cases.

Spelling out ‘closure for easy cases’ isn’t straightforward. Suppose we did it as follows:

> (10.1) If an agent knows that $A$ and that $A$ implies $B$, and there’s an easy argument from ‘$A$’ to ‘$B$’, then she knows that $B$.

The problem here is that easiness isn’t transitive. There may be an easy argument from $A_1$ to $A_2$, from $A_2$ to $A_3$, and so on, right through to some $A_n$. Chaining all these arguments gives us an argument from $A_1$ to $A_n$; but this may not itself be an easy argument. But if easy arguments bring closure in their wake, then so will the long, difficult arguments we get by chaining them. Indeed, if we restrict ‘argument’ to ‘deductively valid argument’, every argument is a chain of easy arguments, and so ‘easy’ closure entails full closure. So this attempt to limit closure to ‘easy’ cases hasn’t got us anywhere.

Here’s a better way to capture a restricted version of closure:

> (10.2) If an agent knows that $A$ and that $A \rightarrow B$, then she is in a position to know that $B$. If she then competently deduces and hence comes to believe that $B$ on this basis (whilst retaining what she knows), she thereby knows that $B$.

Hawthorne (2005, 29) and Williamson (2000, 117) endorse a similar epistemic closure principle. We find this weak closure principle plausible. It says that competent deduction preserves knowledge. It’s hard to see how things could be otherwise. This principle, weak as it is, is strong enough for scepticism to bite. For suppose you competently deduce as above:

> (10.3) I’m reading *Impossible Worlds*.
>
> (10.4) If I’m reading *Impossible Worlds*, then I’m not in a sceptical scenario.
>
> (10.5) Therefore, I’m not in a sceptical scenario.

**(p.221)** Yet, it seems, you don’t know that conclusion. Then you can’t have known both premises. But (10.4) is a priori and easily established; it’s hard to see how you could fail to know it, given that (as a rational person) you believe things to be so. Conclusion: you don’t know (10.3). Scepticism prevails.

We’re not arguing for scepticism. We think we know stuff. Our point is merely that the impossible worlds approach as such doesn’t help with the deep philosophical issue of scepticism. Indeed, we’d be dubious of any formal

semantics that claimed to do better. Scepticism is a deep philosophical problem, which requires a philosophical solution.

## 10.3 The Problem of Bounded Rationality

Any attempt to deal with the knowledge and beliefs of rational but non-ideal agents faces a deep problem. What must they know (or believe), given what else they know (or believe)? This is the issue of the *granularity* of epistemic and doxastic states, which we encountered in §8.4. There, we argued that there are no non-trivial closure conditions on epistemic states. So we reject closure conditions (C1–C2) and (C4–C8) for knowledge, and (C1)–(C8) for belief. ((C3) is valid for knowledge simply because knowledge is factive, which guarantees that one cannot know contradictory things.) This ensures that epistemically possible worlds include logically impossible worlds, some of which are not closed under *modus ponens*.

We also accept a weak epistemic closure principle, (10.2). But this in itself tells us nothing directly about the nature of epistemic possible worlds. It tells us that certain agents are *in a position* to know something, not that they do in fact know it. That's compatible with epistemically possible worlds obeying no purely epistemic closure principles whatever. (Rather, we should capture (10.2) in terms of a constraint on accessibility relations: doxastic accessibility must align with epistemic accessibility to the extent required by (10.2).)

**(p.222)** The difficulty we find in accounting for rational yet non-ideal agents is that, from the theorist's point of view, the *rational* component seems in deep conflict with the *non-ideal* component. Since such agents are non-ideal reasoners, they don't know all that follows from what else they know. Yet since they are rational, the 'anything goes' approach, on which knowing something does not imply knowing anything else in particular, would seem inappropriate.

It's sometimes said that rational but non-ideal agents know whatever follows *easily* from what they know. Chalmers (2010), for example, says that

> it is plausible that $p$ is epistemically possible when one could not *easily* come to know that $\neg p$ given what one already knows. The corresponding notion of deep epistemic possibility is something like the following: it is deeply epistemically possible that $p$ when $\neg p$ is not easily knowable a priori.
>
> (Chalmers 2010, 105–6)

This is a natural thought. But teasing out the details is not at all simple. Suppose that one could easily come to know that $A$, given what one knows. Then, given Chalmers's suggestion, $\neg A$ isn't epistemically possible. If this implies that $A$ is epistemically necessary (for that agent, given what she already knows), it immediately follows that she knows that $A$. This gives us the 'easy closure'

principle, (10.1) from §10.2: the agent knows whatever follows easily from what she knows. Yet this can't be right, as we saw in §10.2, for easiness isn't transitive. Accepting (10.1) forces us to treat our agent as knowing whatever follows deductively from what she knows, given that any deductive reasoning can be broken down into discrete easy chunks.

Any closure principle on knowledge is vulnerable to this kind of reasoning. That's the nature of closure principles: they are all-or-nothing beasts. But 'all' is too much closure, and 'nothing' is too little rationality. This is what we will call (following Jago (2014a)) the *problem of bounded rationality*: the conflict between normative principles of rationality which govern concepts like belief, and our limited cognitive resources.

Let's take a moment to review just what rationality seems to require of belief states. Why is it, exactly, that the 'anything goes' **(p.223)** approach seems inadequate? Davidson (1985), Dennett (1987), Stalnaker (1984), and many others emphasize how belief ascription is a normative practice, whose purpose is ultimately to make rational sense of action. To ascribe a belief, we must first treat the agent in question as a rational being. It makes no sense to attribute, say, beliefs that $A$ and that $B$, without thereby (perhaps implicitly) ascribing the belief that $A \wedge B$. Moreover, on this view, there's no question of what a belief is, outside this normative practice. As a consequence, belief simply cannot be as fine-grained as the 'anything goes' approach allows. The deep objection to the 'anything goes' approach to belief (and cognate concepts), therefore, is that the analysis fails to capture its essential rational basis.

The problem isn't limited to analysing belief. The *problem of rational knowledge* (Jago 2014b) is that the following platitudes are incompatible:

(i) Rational agents seemingly know the trivial consequences of what they know,

(ii) Rational agents do not know all logical consequences of what they know.

Here, we're using 'trivial' (as we did in in §9.5) in a way that includes all the basic inference steps (such as an instance of *modus ponens* or Disjunction Introduction), but which doesn't include all valid inferences, many of which are highly non-trivial.

The problem is that 'trivial', like 'easy', isn't transitive. Closure under trivial consequence brings full closure in its wake. So (i), interpreted as a closure principle, directly conflicts with (ii). Note that, in this formulation, the logical omniscience problem is just one half (ii) of the problem of rational knowledge.

The deep problem is to avoid logical omniscience *without losing sight of the agent's rationality*.

Bjerring (2010, 2012) argues that no solution to these problems is possible. He teases out requirements that epistemic space, the space of all epistemically possible worlds, should satisfy. He then shows **(p.224)** that they are jointly inconsistent. We can set out the essence of his argument as follows. Take any world. If it fails some trivial inference, such as representing both $A$ and $B$ but not $A \wedge B$, then it is blatantly not a way the world might be. Even rational agents with very limited resources recognize this. So such worlds are not epistemic possibilities for any agent. They should play no role in epistemic space. It follows that all worlds in epistemic space, the *deeply* epistemically possible worlds, are closed under all trivial inferences. But, as we've already seen (§9.5), closure under trivial deductive consequence generates full deductive closure.

An adequate response to the problem must argue that such worlds are (deep) epistemic possibilities, even though they are not closed under trivial (or easy) inference. Our approach (as in Jago (2014a)) is to distinguish sharply between what a world represents as *not being the case*, on the one hand, and what the world does not represent as being the case, on the other. Ordinary possible worlds are maximally consistent, and so do not distinguish between not representing something as being the case and representing something as not being the case. Our worlds, however, can be incomplete. So, to represent $A$ as not being the case, a world must explicitly say *it is not the case that* $A$, or $\neg A$. Contrast this with a world which is silent on $A$: it says neither that $A$, nor that $\neg A$.

Following Jago (2014a), we argue that whether a world is deeply epistemically possible depends only on what it represents. Worlds are debarred from being deep epistemic possibilities when they represent a blatant impossibility as being the case. Blatant impossibilities include blatant contradictions. So, a world which represents that $A \wedge \neg A$, or which represents both that $A$ and that $\neg A$, is ruled out. It is not deeply epistemically possible: it cannot be treated as an epistemic possibility, for any possible agent. Similarly, if an agent can easily infer $A \wedge \neg A$ from what a world represents, then that world is ruled out. (We discussed one way to make sense of this idea in §9.5. We'll return to the idea in §10.4.)

Contrast this situation with the case in which one can easily infer some $A$ from what a world represents, where that world doesn't **(p.225)** itself represent that $A$. Such worlds are not ruled out from contention as deep epistemic possibilities, we claim, for they do not explicitly deny $A$. They say nothing about $A$ explicitly (even though $A$ follows easily from what they do say). There is no tension between what they explicitly affirm and what they deny. (This doesn't affect our

analysis of knowledge and belief as kinds of necessity, reflecting what's true according to all worlds accessible to an agent.)

To motivate the idea, consider how every story you've ever heard is *partial* in what it explicitly represents. For any story, there are some facts it remains completely silent on. (We'll say more on how fictions work in Chapter 11.) We're never told what Holmes had for breakfast the day he first met Watson. Either he had toast or he didn't. Yet the story doesn't say he had toast, and doesn't say he didn't. The story is partial. We don't, on that basis, reject the *Holmes*-stories as epistemic impossibilities. Rather, we treat them as partial representations of what we take to be determinate and complete states of affairs. That's how we should think of incomplete worlds when we're assessing whether they count as epistemic possibilities. So we should allow that they're deeply epistemically possible, just so long as they don't explicitly represent some blatant impossibility.

## 10.4 Bounded Rationality and Vagueness

We have to admit, we find puzzling the picture of bounded rationality we've described. We've argued (following Jago 2014a) that it's legitimate to treat incomplete worlds as epistemic possibilities. But as a result, we will have epistemic states which do not capture *trivial* logical truths, which can easily be recognized as valid. Incomplete worlds need not represent trivial logical truths, such as $A \vee \neg A$. Having accepted such worlds as epistemic possibilities, we find ourselves with agents who do not believe $A \vee \neg A$. How can we call such agents *rational*?

This is the problem of bounded rationality emerging again. An agent who does not believe all consequences of what she believes **(p.226)** must fail to believe some trivial consequence of what she believes. But we cannot *say* that she fails to believe that consequence, without thereby treating her as being irrational. (And perhaps, in that case, we should not describe her cognitive state in terms of belief at all.)

There is a parallel to the literature on *vagueness* here. (In what follows, we draw on Jago 2014b.) Consider the essay marking example from §9.5, in which each of 100 essays, ranging from excellent to dire, gets a different mark. It seems absurd to pronounce for sure that only those essays scoring (say) 55% or more were any good. Was the 54% 'not good' essay that much worse than the 'good' one scoring just 1% more? Yet we can't capture what's wrong with this by saying: if one essay was good, then so was the one scoring just 1% less. That principle entails that all or none of the essays were good, and that's clearly wrong.

One response to the problem is that, although there is a fact out there about precisely which essays were good, we can't pinpoint that fact with any precision (Williamson 1994). We just can't know enough about how 'good essay' is used to

determine precisely which essays were the good ones. And since we shouldn't assert what we don't know (according to Williamson (1996, 2000)), we can't ever be in a position to assert things like 'only the essays scoring over 55% were any good'. (DeRose (1992), Hawthorne (2004), Stanley (2005), and Schaffer (2008) all support this *knowledge norm* for assertion.)

Jago (2014b) argues that something similar happens in the case of knowledge and belief ascriptions. A non-ideal agent will fail to know (or believe) some trivial consequence of what else she knows (or believes). This is an *epistemic oversight*: a particular case in which the agent fails to know (or believe) a particular trivial consequence of what she knows (or believes). But knowledge and belief ascriptions are part of a normative practice of explaining behaviour. We cannot be in a position to know, and hence can never assert, just which trivial consequence an agent fails to know (or believe). Epistemic oversights exist, but we can never put a finger on them.

The considerations here are very similar to those we met in §9.5. The idea pursued there was that the content of valid deductions is **(p.227)** indeterminate, because it may be indeterminate whether a world $w$ is deeply epistemically possible. The related idea here is that states of belief and knowledge are themselves vague, because it's indeterminate which logical consequences of her beliefs an agent believes. These ideas are deeply connected. If it is indeterminate whether an inference is informative, then it may be indeterminate whether an agent who determinately believes its premises thereby believes its conclusion.

Belief (and knowledge) states are themselves vague. It's indeterminate what an agent believes (or knows), given what else she believes (or knows). It might be determinate that our agent believes that $A$, but indeterminate whether she believes that $B$, for some related $B$. Yet there must be constrains on the extent of this indeterminacy. Something like the following must hold:

> (TRIV) If it's determinate that an agent knows that $A$, and the inference from '$A$' to '$B$' is trivial, then she cannot determinately fail to know that $B$.

A principle along the lines of TRIV brings with it considerable explanatory power, if we also accept either knowledge or determinate truth as the norm of assertion (Williamson 1994). Suppose that one may assert only what is determinately true. (This is implied by taking knowledge as the norm of assertion, since one cannot know what's indeterminate.) Then one can assert that agent $x$ believes that $A$, but not $B$, only if it is determinate both that $x$ believes that $A$ and that she does not believe that $B$. But, if the inference from '$A$' to '$B$' is trivial, this is precisely the situation ruled out by TRIV. So, when '$A$' trivially entails '$B$', we can never assert that an agent believes that $A$ but not that

*B*. In general, we can never assert any failure of trivial closure in an agent's belief state.

This feature, we think, is what gives the misleading impression that belief states are closed under trivial consequence. As we have seen, they cannot be closed under trivial consequence. Yet we can never discern or assert any counter-instance. We mistakenly go from our inability to falsify the closure principle through counter-examples to its truth.

## **(p.228)** 10.5 Belief and Trivial Inference

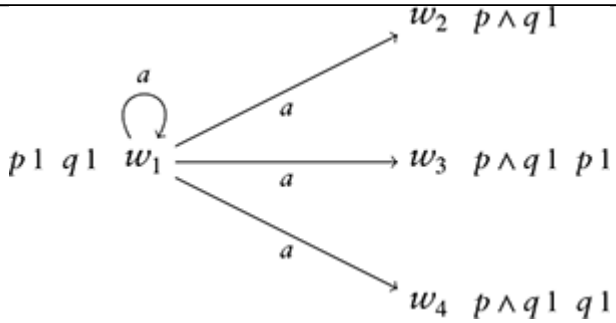
In this section, we'll provide formal models of belief (and knowledge), following Jago (2014b), which capture the idea that belief (and knowledge) states are themselves vague (§10.4). We'll then show how TRIV from §10.4 falls out of these models. This section is largely a technical exercise in showing how precise formal models can validate TRIV. We follow the general approach of Chapter 5, by imposing doxastic and epistemic accessibility relations on a domain of epistemic scenarios. Our domain is an *epistemic space* (§9.5), consisting of deep epistemically possible worlds. To model belief, we add a doxastic accessibility relation between worlds in the space for each agent under consideration. We can model knowledge by imposing further conditions (including reflexivity) on these relations. (Reflexivity ensures that knowledge implies truth.) And we can model both knowledge and belief by adding a doxastic and an epistemic accessbility relation for each agent, restricted so that the former implies the latter. (This ensures that knowledge implies belief.) To keep the presentation simple, we'll focus on models with doxastic accessibility relations only.

As we saw in §9.5, *epistemic space* is a vague notion. It is indeterminate just which impossible worlds count as deep epistemic possibilities, and hence indeterminate which worlds make up epistemic space. The make-up of epistemic space is governed by our analysis of deep epistemic possibility (§9.5):

> (EP) World *w* is (deeply) epistemically possible just in case *w* isn't the root of any small world-proof.

A world-proof is a structure imposed by reinterpreting sequent calculus rules as relations between worlds. Its size reflects how difficult it is (in terms of number of rule-applications) to uncover a hidden contradiction within a world.

If it is indeterminate whether a world is deeply epistemically possible, then it may be indeterminate whether that world is **(p.229)** epistemically accessible for any agent. As a consequence, belief states may be indeterminate (just as the information contents of §9.5 may be indeterminate). Nevertheless, if we impose accessibility relations on epistemic space in the ordinary way, then many facts about an agent's belief will be determinate. As a very simple example, consider the following model, with $w_1$ the only possible world:




Here, $w_1$ is a possible world representing that $p$ and that $q$, and hence that $p \wedge q$. By contrast, $w_2$, $w_3$, and $w_4$ are all incomplete but consistent worlds, representing just $p \wedge q$, just $p \wedge q$ and $p$, and just $p \wedge q$ and $q$, respectively.

In this model, agent $a$ at $w_1$ believes that $p \wedge q$, but not that $p$ or that $q$. Since both $w_1$ and $w_2$ are consistent, neither are associated with a world proof, and so, determinately, both are deep epistemic possibilities. (We argued in §10.3 that this is reasonable, for an agent may view $w_2$ as an incomplete description of a possible world.) So, it's determinate that our agent believes that $p \wedge q$ and determinate that she doesn't believe that $p$. Then there should be no problem with our asserting that she believes that $p \wedge q$ but not that $p$.

This is deeply problematic. In §10.4, we argued that there must be epistemic oversights: cases in which the agent fails to believe trivial consequences of what she believe. But we also claimed that there are no determinate cases of an epistemic oversight. They exist, but we can never put our finger on one. We are never in a position to say that our agent's belief state gives out at *this* particular point. We suggested the TRIV principle:

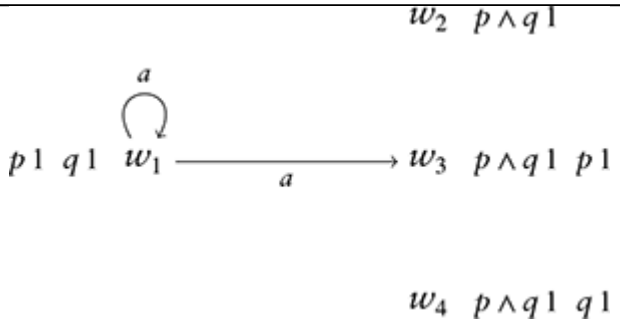
**(p.230)**

> (TRIV) If it's determinate that an agent knows that $A$, and the inference from '$A$' to '$B$' is trivial, then she cannot determinately fail to know that $B$.

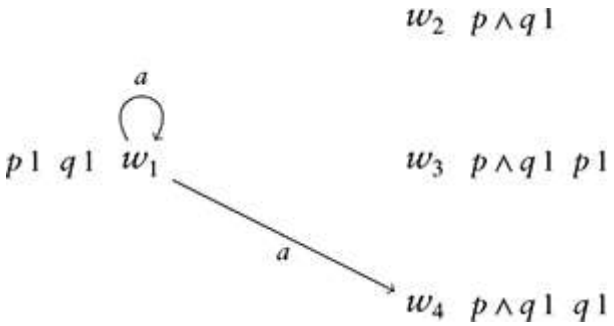
Our simple model above invalidates this principle. To capture it, epistemic models require a little more structure.

The suggestion in Jago 2014b is that an agent's epistemic accessibility relation itself is indeterminate, and the indeterminacy is not merely due to the indeterminacy of epistemic space. In our model above, all four worlds are eligible epistemic possibilities. The suggestion is that a highly incomplete word like $w_2$ can never be determinately accessible for an agent. A more complete version of $w_2$ might represent that $p$ (as $w_3$ does) or that $q$ (as $w_4$ does). A more complete version still would represent both that $p$ and that $q$ (as $w_1$ does).

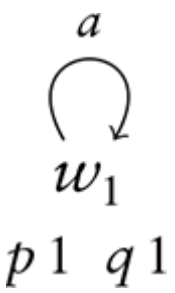
One alternative version of $a$'s epistemic accessibility relation considers only worlds which are explicit about $p$, as in this model:




Another alternative considers only worlds which are explicit about $q$, as in this model:




**(p.231)** What's determinate is whatever holds relative to all of these alternatives. The determinate accessibility relations, for example, are those that hold in all alternative models: from $w_1$ to $w_1$, in this case. But this doesn't mean that our model shrinks to $w_1$:




In this single-world model, agent $a$ is logically omniscient (since $w_1$ is a possible world). On our approach, by contrast, it's *determinate* that some incomplete world is epistemically accessible. That guarantees that the agent's belief state isn't logically closed. But it's indeterminate which incomplete worlds are accessible, and so indeterminate where the failure of closure occurs.

(How we interpret this multiplicity of models, philosophically, is up for grabs. One option is to say that one model gets things right, but we can't ever know which. This is the *epistemicist* approach to vagueness (Williamson 1994). Another option is to say that no one model gets things right, for each model is precise in a way that reality is not. Rather, what's true amounts to what is the case according to all the models (and so equates to what's determinately true). This is the *supervaluationist* approach to vagueness (Fine 1975b). Both

approaches to vagueness have their problems, of course. The aim here isn't to solve the problem of vagueness, but to bring attempted solutions to that general problem to bear on the specific problem of epistemic oversights.)

Now for the formal details, which we give for multiple agents. We use our standard propositional language $\mathscr{L}$ with knowledge operators $K_i$ for each agent $i$ and an operator '$\triangle$', read is 'it is determinately the case that …'. We define '$\triangledown A =_{\text{df}} \neg\triangle\neg A$', read as 'it is indeterminate whether …'. These operators allow our object language to express matters which are (or are not) vague. For convenience, we'll phrase our semantics for the '$K_i$'s in terms of epistemic projection functions (which we encountered in §5.1).

**(p.232)** Definition 10.1 **(Epistemic models)** *An epistemic model for k agents is a tuple $M = \langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle$, with W, N, and ρ as in definition 9.1, and each $f_i$ is the epistemic projection function for agent i, assigning a subset of W to each $w \in W$. As before, a pointed model is a pair, $\langle\!\langle M, w \rangle\!\rangle$ where w is a world in M. We abbreviate $\langle\!\langle M, w \rangle\!\rangle$ to $M^w$. The rank of $M^w$ is #w, as given by definition 9.2.*

Given the precise projection functions $f_1, \ldots, f_k$, we then define the alternative projection functions, one for each sentence $A$ for each agent $i$, as follows.

Definition 10.2 **(A-variant of $f_i$)** *Given a model M as above, we set:*

$$f_i^A w = \begin{cases} (f_i w \cap \{w \mid A \in V^+ w\}) & \text{if } f_i w \subseteq \{w \mid A \notin V^- w\} \\ \qquad\qquad \cup (f_i w \cap W^P) & \\ f_i w & \text{otherwise} \end{cases}$$

*Let $f_i^{\mathscr{L}} = \{f_i\} \cup \{f_i^A \mid A \in \mathscr{L}\}$.*

Definition 10.3 **(Alternative sequences)** *For an epistemic model M for k agents as above, let $\mathbf{A}_M = \{ \langle\!\langle\, g_1 \cdots g_k \,\rangle\!\rangle \mid g_i \in f_i^{\mathscr{L}}, i \leq k\}$. If $\alpha \in \mathbf{A}_M$ is an alternative sequence, then $\alpha^i$ (for $i \leq k$) denotes the ith member of α, i.e., an alternative projection function for agent i.*

Next, we define a notion of truth for the whole language, relative to an alternative sequence. We extend the '$\rho$' notation, writing '$\rho_w^\alpha A1$' to mean that $A$ is true at $w$, relative to alternative sequence $\alpha$.

Definition 10.4 **(α-truth)** *Given an epistemic model M as above, we define $\rho_w^\alpha$ as follows. For possible worlds $w \in N$ and where A is an atom, negation, conjunction, disjunction, or material implication, we set $\rho_w^\alpha A1$ iff $\rho_w A1$, with $\rho_w$ as in definition 9.1. For the remaining cases (with $w \in N$), we set:*

(SK) $\rho_w^\alpha(K_i A)1$ *iff* $\rho_{w_1} A1$ *for all* $w_1 \in \alpha^i w$

(S△) $\rho_w^\alpha(\triangle A)1$ *iff* $\rho_w^\beta A1$ *for all* $\beta \in \mathbf{A}_M$

**(p.233)** *In all these cases, we set* $\rho_w^\alpha \triangle A0$ *iff not* $\rho_w^\alpha A1$. *For impossible worlds* $w \in W - N$, *we set* $\rho_w^\alpha A1$ *iff* $\rho_w A1$ *and* $\rho_w^\alpha A0$.

Definition 10.5 **(*n*-entailment)** *Given a pointed epistemic model* $M^w$ *where* $M = \langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle$, *A is true in* $M^w$ *iff* $\rho_w^{\langle\!\langle f_1 \cdots f_k \rangle\!\rangle} A1$. *Then, for any* $n \in \mathbb{Z}^+ \cup \{\omega\}$, *logical n-entailment,* $\models_n$, *is defined as:*

$\Gamma \models_n A$ *iff, in every pointed model* $M^w$ *of rank* $r \geq n$ *where* $w \in N$, *each* $B \in \Gamma$ *is true in* $M^w$ *only if A is true in* $M^w$.

It is then easy to see that $\models_n$ extends classical (propositional) entailment:

**Theorem 10.1** *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$: *if* $\Gamma$ *classically entails A, then* $\Gamma \models_n A$.

*Proof:* By contraposition. If $\Gamma \not\models^n A$, then for some pointed model $\langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle^w$ of rank $r \geq n$ with $w \in N$, each $B \in \Gamma$ is true but $A$ is not. Set $VA = 1$ iff $\rho_w^{\langle\!\langle f_1 \cdots f_k \rangle\!\rangle} A1$. Then it is easy to see that $V$ is a classical valuation function over atoms $\{p, \triangle A, K_i A \mid p, A \in \mathscr{L}\}$. Since $VB = 1$ for each $B \in \Gamma$ but $VA = 0$, it follows that $\Gamma$ does not classically entail $A$.

Now recall the formal notion of *n-trivial consequence*, $\mathrm{triv}_n(\Gamma, A)$, from §9.5 (definition 9.3). Both of our key formal concepts, *n*-entailment and *n*-trivial inference, are parameterized by an integer (or $\omega$) $n$, which we think of as an artificial precisification of vague epistemic space. Given any such precisification, we can show that our models imply TRIV: if it's determinate that an agent knows each of the premises $\Gamma$ of a trivial inference from $\Gamma$ to $A$, then it's not determinate that she fails to know that $A$.

**Theorem 10.2** *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$, *if* $\mathrm{triv}_n(\Gamma, A)$ *then* $\{\triangle K_i B \mid B \in \Gamma\} \models_n \neg\triangle\neg K_i A$.

**Corollary 10.3** *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$, *if* $\mathrm{triv}_n(\Gamma, A)$ *then* $\{\triangle K_i B \mid B \in \Gamma\} \cup \{\neg K_i A\} \models_n \triangle K_i A$.

**(p.234)**

**Corollary 10.4** *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$, *if* $n \geq 3$, *then* $\models_n \neg\triangle\neg K_i(A \vee \neg A)$ *and* $\neg K_i(A \vee \neg A) \models_n \nabla K_i(A \vee \neg A)$.

*Proof sketch:* Suppose that $\mathrm{triv}_n(\Gamma, A)$. Then for any pointed epistemic model of rank $r > n$ where $\Gamma$ is true, $A$ is not false. Suppose also that $\triangle K_i B$ for each $B \in \Gamma$ is true in some $M^w$. Then, for each alternative accessibility relation and each $i$-accessible world $w'$, $B$ is true at that world. Then $A$ is

not false at $w'$. But, since the alternative projection function $f_i^B$ forces a stance on $A$, $A$ is true at each world $w' \in f_i^B w$. Thus $K_iA$ is true at $w$ relative to any alternative sequence containing $f_i^B$. Then since $\neg K_iA$ is false at $w$ relative to some alternative sequence, $\triangle\neg K_iA$ is false, hence $\neg\triangle\neg K_iA$ is true, in $M^w$. Corollary 10.3 follows given $\triangledown A =_{\text{df}} \neg\triangle A \wedge \neg\triangle\neg A$, and corollary 10.4 is a special case. Full proofs can be found in Jago 2014b.

However we choose a precise delineation of 'epistemic scenario' and 'trivial consequence', if the inference from $\Gamma$ to '$A$' is trivial then determinate knowledge of $\Gamma$ entails the agent does not determinately fail to know that $A$, just as our principle TRIV from §10.4 says. Equivalently (as corollary 10.3 says), if agent $i$ does not know some trivial consequence '$A$' of what she knows, then it is indeterminate whether she knows that '$A$'. So, on the account proposed, there are no determinate epistemic oversights. Each case of an epistemic oversight is an indeterminate case.

Since what is indeterminate is not rationally assertible, it is then never rational to assert that agent $i$ suffers from a particular epistemic oversight. If an agent is not logically omniscient, then we can be sure that she suffers from some epistemic oversight. Indeed, it is determinate that real-world agents are not logically omniscient, and hence determinate that real-world agents suffer from epistemic oversights. But we can never say what they are: we cannot locate them in a rational agent's epistemic state. Whenever we focus on a particular trivial consequence '$A$' of agent $i$'s knowledge, it is never rational to assert that she does not know that $A$ (even if that's the case). In this way, our formal models support our philosophical contention **(p.235)** that epistemic oversights are always elusive, just as counterexamples to tolerance principles for vague predicates are.

## 10.6 Believing Contradictions

There's a worry that's been building up since §9.5, where we introduced our concept of *epistemic space*. There, we excluded explicitly contradictory worlds (according to which some $A$ is both true and false) from epistemic space. We also excluded worlds from which we can reach explicitly contradictory worlds in a relatively short number of steps. All these excluded worlds are associated with a small world-proof, and are excluded on that basis. Our key principle is this:

(EP) World $w$ is epistemically possible just in case $w$ isn't the root of any small world-proof.

For some worlds, it's indeterminate whether they satisfy the criterion (because it's indeterminate whether any associated world-proof is small). But it's determinate that explicitly contradictory worlds are not epistemic possibilities

(for any agent). So they played no role in our analysis of epistemic and doxastic states (§10.5).

The objection is simple: surely there *are* agents who believe explicit contradictions. Whether they're right or wrong to do so isn't the point, for beliefs needn't be true, or even reasonable. People can believe all sorts of things. There's even a well worked out philosophical view, *dialethism*, according to which contradictions can be true. And yet, we're saying that it's impossible to believe a contradiction. How can we square this?

Take the case of Graham Priest, dialethism's foremost proponent (Priest 1979, 1987, 2014, 2016b). Few people on earth are more rational and logically adept. Fewer still know more about negation. So when Priest says, clearly and repeatedly, that he believes contradictions, and backs this up with sophisticated philosophical and logical argument, how can we disagree with him?

**(p.236)** Dialethists clearly believe something when they assert a contradiction. We follow Jago (2014a, §7.5) in thinking that their assertions mean something a little different from what they would seem to mean. In particular, the dialethist may mean something a little different from what other English speakers typically mean by 'not'. The evidence for this comes out in the logical rules she takes to govern her concept of negation. In classical and intuitionistic logic, the logical rules for 'not' allow us to link assertions to denials (or what we accept to what we reject).

Suppose in a conversation we keep track of the things we accept, $\Gamma$, and the things we reject, $\Delta$, by writing: $\Gamma \vdash \Delta$. Then, if we accept $A$, we should reject $\neg A$, and vice versa:

$$\frac{\Gamma, A \vdash \Delta}{\Gamma \vdash \neg A, \Delta} \qquad \frac{\Gamma \vdash A, \Delta}{\Gamma, \neg A \vdash \Delta}$$

These are the sequent rules for classical negation. The dialethist rejects them both. Her concept of negation differs from the classical one. (She readily agrees, for she finds the classical notion incoherent.) The question, then, is what is typically expressed in English by 'not'.

The line in Jago 2014a, §7.5, then goes as follows. In communicating with each other, we need some way to indicate what we accept and what we reject (or deny). We could have evolved green and red lights on our heads, so that uttering '$A$' with a green light amounts to accepting that $A$, whereas uttering '$A$' with a red light amounts to denying it. Or we could put a thumb up or down as we utter '$A$'. But there's a simpler way: we use 'no', 'not', and cognates, to signal disagreement and denial. (That's a very useful tool, since we can then deny a denial, by using 'not' twice.)

This argument is part conceptual, part empirical. The conceptual part is that, to engage in the kind of discursive practices our complex interactions require of us, we need a mechanism for signalling acceptance and rejection of contents. The empirical part is that English speakers typically use ‘not’ and ‘no’ for that purpose. But we’re free to step out of that practice. Dialethist uses of ‘not’ may **(p.237)** well express their non-classical, non-explosive concept of negation. They believe those contents, which they express using ‘not’ and we express using ‘dialethist negation’. But those contents are not literally contradictions, because contradictions are those involving *negation*, the concept expressed in standard English by ‘not’. Dialethists believe what they say when they say, for example, ‘the Liar is both true and not true’. Their assertions are contentful and meaningful. But they don’t believe that the Liar is both true and not true.

### Chapter Summary

We outlined the case for making belief states the primary focus of our analysis (§10.1), and for including impossible (as well as possible) worlds in that analysis (§10.2). This allows us to deny various closure principles, although this probably won’t help defeat worries about external-world scepticism (§10.2).

The issue that concerned us most is the *problem of bounded rationality* (§10.3): belief states seem to be closed under ‘easy’ trivial consequence, but not under full logical consequence, and yet the former implies the latter. Our solution was that some trivial closure principle must fail on a given belief state, yet it is indeterminate just where this occurs (§10.4). We cannot know, or be in a position that entitles us to assert, which trivial consequence of her beliefs an agent fails to believe. We gave formal models of belief states alone these lines, and showed that they respect the indeterminacy-of-closure intuition, in §10.5. Finally, in §10.6, we discussed how we might square this approach, which says that no one can genuinely believe a contradiction, with the fact that some people seem to believe just that. **(p.238)**