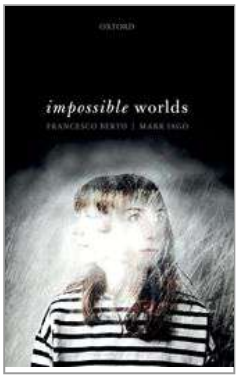

Impossible Worlds
Francesco Berto and Mark Jago

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

Epistemic Logics

Francesco Berto
Mark Jago

DOI:10.1093/oso/9780198812791.003.0005

Abstract and Keywords

Standard possible-worlds epistemic logic gives rise to the problem of logical omniscience. There are attempts to deal with the problem without using impossible worlds. A number of these approaches are discussed in this chapter and all are found wanting. The impossible worlds approach is immediately more successful, but faces a deep problem: how should impossible worlds be constrained, so as to give adequate models of knowledge and belief? One option is to take impossible worlds to be closed under some weaker-than-classical logic. But this approach does not genuinely solve the problem of logical omniscience. A different approach is the *dynamic* one, whereby epistemic states are not closed at any one time, but nevertheless evolve towards closure in a dynamic way.

Keywords: knowledge, belief, logical omniscience, weaker-than-classical logic, epistemic states, dynamic epistemic states

5.1 Standard Epistemic Logic and Logical Omniscience

In §1.2, we introduced the idea of understanding knowledge and belief as restricted quantifiers over possible worlds, where the accessible worlds are those that represent epistemic possibilities for a cognitive agent. This can be modelled by taking the language $\mathscr{L}$ of §4.1 with its normal Kripke semantics, and interpreting '$\Box$' as an operator representing knowledge or belief. We'll rewrite '$\Box$' as 'K', for 'one knows that'. (We'll also talk about belief, and sometimes use 'B' in place of 'K'. But most of what we say about modelling knowledge goes for belief, and vice versa.) Its semantic clause is:

(SK) $v_w(KA) = 1$ if for all $w_1 \in W$ such that Rww_1, $v_{w_1}(A) = 1$, and 0 otherwise.

One can then read the dual '$\Diamond$' as 'it is compatible with what one knows/believes that'.

The relation R in a Kripke model $\langle\!\langle W, R, v \rangle\!\rangle$ should now be read in an epistemic way, as *epistemic accessibility*. Epistemic operators are often indexed to a particular agent, with 'K_iA' read as 'agent i knows that A'. Indexing allows a multi-modal logic representing the cognitive states of a plurality of agents. For agents 1, …, n, each gets its own knowledge operator K_1, …, K_n, with a corresponding epistemic accessibility relation, R_1, …, R_n.

(p.108) It's sometimes useful to rephrase accessibility in terms of a function f which, given a world w as input, returns the set of worlds fw that are epistemically accessible from w. This f is called the *epistemic projection function*, and is defined by setting $fw = \{w_1 \mid Rww_1\}$. Then KA's truth at w requires A's truth at all worlds in fw. When we have multiple epistemic accessibility relations R_1, …, R_n, we have multiple epistemic projection functions, f_1, …, f_n, one for each agent. Since the issues we'll go on to discuss arise in both the single-agent and multi-agent settings, we'll focus for simplicity on single-agent models, with a single modality 'K'.

Our concern here will be with the various issues arising under the heading of *logical omniscience* (§1.3). The issues are the result of our semantics satisfying various *closure conditions*. These take the form: if an agent knows —, she must also know —. Fagin et al. (1995, 335–6) and Van Ditmarsch et al. (2008, 23) discuss the following (with '$A \dashv\models B$' meaning that $A \models B$ and $B \models A$):

(C1) If KA and $A \models B$, then KB

(C2) If $\models A$, then KA

(C3) $\models \neg(KA \wedge K\neg A)$

(C4) If KA, and $A \dashv\models B$, then KB

(C5) If KA and $\models A \supset B$, then KB

(C6) If KA and $K(A \supset B)$, then KB

(C7) If $K(A \wedge B)$, then KA and KB

(C8) If KA, then $K(A \vee B)$

(C1) is *Closure under entailment* or *Full omniscience*. (C2), *Knowledge of all valid formulas*, is a special case of (C1). We met both principles in §1.3. (C3),

Consistency, says that one cannot know contradictory things. (C4), *Closure under logical equivalence*, follows from (C1) as logical equivalence is defined as two-way entailment. **(p.109)** (C5), *Closure under valid implication*, is equivalent to (C1) in systems in which $\vDash A \supset B$ if and only if $A \vDash B$. (C6) is *Closure under known implication*. (C7) and (C8) are often called *Closure under conjunction* and *Closure under disjunction*, respectively. There are corresponding closure principles for belief.

In §1.3, we discussed how (C1) and (C2) are implausible for the case of real, finite, and fallible cognitive agents. The corresponding principles for belief are just as implausible, as is (C3): we are all inconsistent believers. The other principles look no more promising. (C4) and (C5) obviously are in no better position than (C1). (C6) is a hotly contested principle in epistemology, because of its relation to external-world scepticism. You know that, if you have hands, then you're not a brain in a vat. But you don't know that you're not a brain in a vat. (How could you?) Given (C6), it follows that you don't know that you have hands (and similarly for other simple bits of external world knowledge). (Dretske (2005), Holliday (2015), Nozick (1981), and Roush (2010) discuss the issue.) (C7) is perhaps the most plausible principle in the list. As for (C8), it seems implausible that, just because one knows that A, one automatically knows the disjunction of it and any arbitrary B. One may lack the very concepts involved in B.

All of these principles except (C3) hold in the weakest normal modal logic **K** from §4.1, with no conditions on the accessibility relation R. ((C2) is related to (N) and (C6) to the K-principle, for example. (C3) requires Seriality to be valid.) This fact tells us that tampering with the accessibility relation is not going to help us avoid all of these principles. So if we want to understand knowledge and belief in terms of modal logic, we should not work with a normal modal logic.

Non-omniscience is often taken as evidence that knowledge and belief are hyperintensional. There are a number of ways to draw distinctions more fine-grained than those available in standard possible worlds semantics, however. Not all of those options resort to impossible worlds. We'll briefly discuss some of them in §5.2.

(p.110) 5.2 Dealing with Omniscience without Impossible Worlds

One radical way to avoid Logical Omniscience is to adopt a *syntactic approach* to modelling knowledge and belief. What an agent knows (or believes) is captured as set of sentences, whose content reflects what the agent knows (or believes). Alechina and Logan (2002), des Rivieres and Levesque (1686), Eberle (1974), Konolige (1986), Moore and Hendrix (1979), and Morreau and Kraus (1998) all take an approach along these lines. Philosophical motivation for their approach may be found in Quine, who sees propositional attitudes as 'involving something

like quotation of one's own imagined verbal response to an imagined situation' (Quine 1960, 219).

Syntactic models of knowledge include a database $\mathscr{D}$ of formulas, and take KA to hold iff $A \in \mathscr{D}$. $\mathscr{D}$ is logically unstructured. It is merely a set of formulas, having no non-trivial logical closure features. As a result, all closure conditions for K are destroyed as well. But the solution seems cheap: knowledge or belief in syntactic structures have few interesting features. As Fagin et al. (1995) say,

> One gains very little intuition about knowledge from studying syntactic structures … In these approaches knowledge is a primitive construct. … Arguably, these approaches give us ways of *representing* knowledge rather than *modelling* knowledge. In contrast, the semantics given to knowledge in Kripke structures *explains* knowledge as truth in all possible worlds.
>
> (Fagin et al. 1995, 345)

An alternative approach draws on Scott-Montague *neighbourhood semantics* (Scott 1970), and represents what an agent knows or believes as an unstructured set $\mathfrak{P}$ of propositions, rather than formulas. Propositions here are understood as sets of possible worlds. $\mathfrak{P}$ is then an unstructured set of sets of worlds. This approach invalidates various forms of logical omniscience.

One that remains valid is (C4), though: if A and B are equivalent and KA holds, then KB holds as well. In a possible worlds setting, **(p.111)** *that* $2 + 2 = 4$ and *that* $x^n + y^n = z^n$ *has no solutions in integers for* $n > 2$ are the same proposition, namely the total set of worlds. Yet one can know the former without knowing the latter.

Some proposals combine a syntactic and a possible worlds approach. *Awareness logic* (Fagin and Halpern 1988) works with three main notions. *Awareness* is syntactic: an agent is aware of A when A belongs to a set of formulas, its 'awareness set'. *Implicit knowledge* gets the standard possible worlds definition, whereas *explicit knowledge* is defined as the combination of implicit knowledge and awareness. An agent explicitly knows that A when she implicitly knows that A and A is in her awareness set. The underlying idea is that lack of omniscience can come from lack of awareness, understood as lack of conception (Schipper 2015). Because explicit belief requires awareness, and awareness is represented via membership of an arbitrary set of formulas, explicit belief can invalidate any non-trivial logical closure condition.

Differentiating between explicit and implicit representational mental states is as such cognitively plausible and independently motivated. The distinction between explicit and implicit memory and knowledge is frequently made in empirical psychology. Explicit memory is often taken to be conscious, involving the deliberate recall of previously acquired information. Implicit memory, by

contrast, is taken to involve a change in performance or in the execution of a task in the light of previously acquired information without conscious recall (Schacter 1986, Schacter and Tulving 1994).

Similarly, representational accounts of belief claim that one believes A explicitly when one has a representation with content A actually present in the mind, as a token of a sentence inscribed in a 'belief box'. One believes A implicitly if one believes A without having a representation with that content present in one's mind (Dennett 1978, 1987). Dennett proposes that, in order for one to believe something implicitly, it is enough that 'the relevant content be swiftly derivable from something one explicitly believes' (Schwitzgebel 2015, §2.2.1).

What does *swift derivability* mean here? In the Fagin-Halpern approach, implicit belief is closed under classical logical consequence. **(p.112)** But this is precisely what spells trouble in the light of logical omniscience. In what sense does a finite cognitive agent have implicit belief in the infinitely many logical consequences of what it believes? This notion of implicit belief (or knowledge) seems more like the implicit rational commitments of one's beliefs (or knowledge). That's not the psychologically motivated notion of implicit belief just discussed (for criticisms along these lines, see (Schipper 2015, 88)).

But now, without a satisfactory notion of implicit knowledge or belief, the awareness approach fares little better than a purely syntactic approach to knowledge and belief representation (Jago 2006, Konolige 1986). Konolige sums up the situation:

> the logic of general awareness represents agents as perfect reasoners, restricted to considering some syntactic class of sentences. There don't seem to be any clear intuitions that this is the case for human or computer agents.
>
> (Konolige 1986, 248)

We now turn to approaches that resort to impossible worlds, with the aim of seeing whether some of them can do better.

5.3 Impossible Worlds for Knowledge and Belief

The idea of adopting impossible worlds in order to address the logical omniscience problem goes back to Hintikka (1975). He proposed that epistemically accessible worlds need not be genuinely logically possible. Instead, he allows the epistemically accessible worlds to be 'options which only look possible but which contain hidden contradictions' (Hintikka 1975, 476).

Rantala (1982a) gives an example of this strategy. Recall how, in Rescher and Brandom's work (§4.4), there are worlds at which conjunction and disjunction behave abnormally. In Rantala's approach, the idea is extended to all logical

operators. Take once again the language $\mathscr{L}$ from §4.1 and give it the following semantics. A *Rantala frame* $\mathscr{F}$ for $\mathscr{L}$ is a triple $\langle\!\langle W, N, R \rangle\!\rangle$, with W the set of worlds, **(p.113)** $N \subseteq W$ the subset of normal, possible worlds, $W - N$ the non-normal or impossible worlds. R is as before. A frame becomes a Rantala model $\mathscr{M} = \langle\!\langle W, N, R, v \rangle\!\rangle$ when endowed with a valuation function v assigning truth values to formulas at worlds. At possible worlds in N, atomic formulas are directly assigned 1 or 0, and compound formulas are evaluated recursively. At impossible worlds in $W - N$, by contrast, *all* formulas are assigned a truth value by v directly, not recursively. Logical consequence and validity are defined, again, as truth preservation at all possible worlds in all models.

As a consequence, at impossible worlds, all formulas are treated as if they are atomic. $A \vee B$ may turn out to be true even though both A and B are false (impossible worlds may be non-prime), and $\neg A$ may turn out to be true when A is (impossible worlds may be inconsistent). This is the generalization of a strategy already met in §4.2: taking impossible worlds as worlds where the logical syntax of formulas can be disregarded when assigning them a truth value. As a consequence, the impossible worlds in $W - N$ are not closed under any consequence relation other than *identity*, $A \vDash A$. Because of this, impossible worlds of this kind are called *open worlds* in Priest (2005).

On this approach, none of of (C1)–(C8) hold. For instance, against (C1) and (C8), consider the following model, with $N = \{w\}$ and the arrow representing accessibility:

At w, Kp is true (since p is true at w_1) but $K(p \vee q)$ is not (since $p \vee q$ is false at w_1). So $Kp \nvDash K(p \vee q)$, even though $p \vDash p \vee q$. (Note that this is no countermodel to $p \vDash p \vee q$. Since $w_1 \in W - N$, it does not affect logical consequence.)

Rantala (1982b) extends this approach to quantified modal logics, and Wansing (1990) develops it into a unified framework for epistemic logics. Wansing shows that various logics for knowledge and belief in Artificial Intelligence, including Fagin and Halpern's **(p.114)** awareness logic (mentioned in §5.2), have equivalent impossible worlds models (that is, models which validate precisely the same formulas). Sillari (2008) establishes further equivalence results in the area.

A naive Rantalian approach, however, faces a serious problem. As noted by Jago (2007, 2009b), the way in which Rantala models manage to invalidate all forms of logical omniscience involves having no restriction on the impossible worlds one can look at, via the epistemic accessibility relation R of the models. The set of worlds W in the frames can include worlds not closed under any non-trivial

consequence relation. Worlds, thus, can correspond to arbitrary sets of formulas of $\mathscr{L}$.

Given world *w* where our epistemic agent is located, then, let $\mathscr{S}_w = \{w_1 \mid Rww_1\}$, the set of worlds accessible from *w*. Let $\mathscr{C} = \{A \mid v_{w_1}(A) = 1 \text{ for all } w_1 \in \mathscr{S}_w\}$, the set of formulas true at all of them. The agent's epistemic or doxastic state can be reduced to a merely syntactic structure: *KA* holds at *w* just in case $A \in \mathscr{C}$, and $\mathscr{C}$ can be a set of formulas lacking any (non-trivial) closure property. The content of epistemic states, then, comes out as highly structured as the syntax of the language. 'Unconstrained' impossible worlds semantics makes no real progress with respect to a merely syntactic approach. One can add constraints on the accessibility relation, or on the logical behaviour of the accessible worlds, which will validate some inferences. As we are about to see, though, how this should be done is no trivial issue.

5.4 Closure under a Weaker Logic

We've seen that, by adding impossible worlds with no logical structure whatsoever, the worlds approach seems no better than the syntactic approach. A natural thought is that we should insist that impossible worlds have *some* degree of logical structure, although not as much as logically possible worlds. Cresswell (1973) and Levesque (1984) **(p.115)** offer approaches along these lines. We'll present a simple version in this section.

Our language $\mathscr{L}$ will include the operators ¬, ∧, ∨, and *K*, with $A \supset B$ being defined as $\neg A \vee B$. As before, a frame $\mathscr{F}$ is a pair 《*W, R*》, with *W* the set of worlds and *R* the epistemic accessibility relation. A frame becomes an *FDE model* $\mathscr{M} = \langle\!\langle\ W, R, \rho\ \rangle\!\rangle$ when endowed with a *valuation relation* *ρ*. (We'll explain the name 'FDE' below.) Unlike the usual valuation function, a valuation relation can connect a formula to more than one truth value at a world: *ρ* can relate the atomic formulas of $\mathscr{L}$ to truth ($\rho_w p1$), falsity, ($\rho_w p0$), both, or neither. We thus get rid of the assumption, embedded in the semantics of classical logic, that truth and falsity are exclusive and exhaustive.

We extend *ρ* to the whole language via the following recursive clauses. We now need to spell out truth and falsity conditions separately for each operator, for now not being true (not being related to 1) is distinct from being false (being related to 0):

(S1¬) $\rho_w(\neg A)1$ iff $\rho_w A0$

(S2¬) $\rho_w(\neg A)0$ iff $\rho_w A1$

(S1∧) $\rho_w(A \wedge B)1$ iff $\rho_w A1$ and $\rho_w A1$

(S2∧) $\rho_w(A \wedge B)0$ iff $\rho_w A0$ or $\rho_w B0$

(S1∨) $\rho_w(A \vee B)1$ iff $\rho_w A1$ or $\rho_w B1$

(S2∨) $\rho_w(A \vee B)0$ iff $\rho_w A0$ and $\rho_w B0$

(S1K) $\rho_w(KA)1$ iff for all $w_1 \in W$ such that Rww_1, $\rho_{w_1} A1$

(S2K) $\rho_w(KA)0$ iff it is not the case that $\rho_w(KA)1$

Logical consequence is truth preservation at all worlds of all models:

$\Gamma \vDash A$ iff for all models $\mathscr{M} = \langle\!\langle\, W, R, \rho \,\rangle\!\rangle$ and all $w \in W$: if $\rho_w B1$ for all $B \in \Gamma$, then $\rho_w A1$

(p.116) (Note that we define logical consequence over all worlds in the model, making no distinction between possible and impossible worlds here.)

This K operator corresponds to what Levesque (1984) calls *explicit* belief. He also defines an implicit belief operator, which is closed under classical logical consequence and hence delivers full logical omniscience (for the notion of implicit belief). We'll focus on the explicit notion only here.

In this semantics, worlds can be inconsistent (making both A and $\neg A$ true, for some A) and incomplete (making neither A nor $\neg A$ true). In the taxonomy of §1.4, they are impossible worlds of the third and fourth kinds: violating classical logic and making contradictions true. Yet they do have some logical structure. They are still *adjunctive*, making $A \wedge B$ true whenever they make both A and B true, and *prime*, making either A or B true whenever they make $A \vee B$ true. They also obey Disjunction Introduction (from A to $A \vee B$) and Double Negation Introduction and Elimination (from A to $\neg\neg A$ and back).

The resulting logic is *paraconsistent*, for $A \wedge \neg A \nvDash B$: contradictions do not entail arbitrary conclusions. It is also *paracomplete*, for $A \nvDash B \vee \neg B$: arbitrary premises do not entail all instances of Excluded Middle. The extensional fragment of this logic (the part lacking K-sentences) is, in fact, one way of presenting *First Degree Entailment*, **FDE** (Belnap 1977, Dunn 1976). This is a simple and well-known paraconsistent and paracomplete logic, and is why we call these structures *FDE models*. We'll also refer to the worlds in those models as *FDE worlds*.

FDE models avoid some problematic forms of logical omniscience. The approach can be used to model agents that have contradictory beliefs (against (C3)), but do not thereby believe everything. Consider the model:

(We display all the atoms related to a truth-value by ρ.) Then $K(p \wedge \neg p)$ holds at w, but Kq does not.

(p.117) Closure under known (or believed) implication (C6) fails too. This is because *modus ponens* fails in the semantics: $A, A \supset B \not\models B$. In the model above, *modus ponens* fails at w_1: $p \supset q$ holds there, for any q, given that $\rho_{w_1} p0$ and that $p \supset q$ is defined as $\neg p \vee q$. For any world that can access such worlds, it may be that KA and $K(A \supset B)$ both hold, and yet KB does not. So in general, $KA, K(A \supset B) \not\models KB$.

Fagin and Halpern (1988) and Jago (2007) stress that logical omniscience strikes back in unwelcome ways, however. Any epistemically accessible FDE world will be adjunctive, closed under Disjunction Introduction, and Double Negation Introduction and Elimination. As a consequence, knowledge and belief come out closed under the corresponding entailments. Conditions (C7) and (C8) still hold.

It is questionable whether this is a good way to model finite and fallible epistemic agents. In particular, one may believe that A without believing that $A \vee B$ for an arbitrary B, for one may lack the B-involving concepts. Similarly, any FDE world making A true also makes true the formula obtained by prefixing an even number of negations to A. But it might be that an agent believes that A, without believing that

$$\overbrace{\neg\neg\cdots\neg\neg}^{\text{1,000,000 times}} A$$

simply because she lacks the cognitive resources to record all the iterations.

More generally, in this system knowledge and belief are closed under weaker-than-classical first-degree entailment. If KA and B is an **FDE**-consequence of A, then KB. But this seems wrong: finite cognitive agents do not know or believe all remote consequences of what they know or believe, even when the notion of entailment in play is **FDE**. There are infinitely many such consequences and they cannot all be computed by a finite mind.

We seem to face a dilemma: either cognitive states like knowledge and belief for real agents are closed under some logical consequence, or they're not. If they are, then logical omniscience returns: we have the implausible situation of an agent that is omniscient with respect **(p.118)** to the target logic. For this not to happen, we seem to have to admit that such states are completely anarchic: they violate any logical closure principle (except for A entailing A). But then, how can we have a *logic* of knowledge and belief at all? Some logicians and AI researchers have conjectured that there is no solution to the dilemma (Meyer and Van der Hoek 1995, 88).

5.5 Going Dynamic

We will now focus on a version of this dilemma, phrased directly in terms of worlds, which we take from Bjerring (2010, 2012). (See also Jago 2014a.) Take a set $\mathscr{R}$ of rules of inference. One may think of a set of sequent calculi or natural deduction rules; but (bracketing issues caused by the presence of rules that discharge temporary assumptions) the point is independent from the specific logical set-up. Call an inference from $A_1, \ldots, A_n$ to B *immediate* when it involves just a single application of a single rule in $\mathscr{R}$.

If what one believes is represented by a set of accessible worlds and these are all closed under the rules in $\mathscr{R}$, then one turns out to be omniscient with respect to $\mathscr{R}$. Suppose, instead, that some accessible world w is not closed under some rule $r \in \mathscr{R}$. Then for some B that follows immediately from some $A_1, \ldots, A_n$, w makes true all of $A_1, \ldots, A_n$ but not B. Then our agent is represented as missing some immediate consequence of what she believes. She considers w as a way things might be, even though it is, in an *obvious* way, not a way things could be. This seems to be a poor approach to modelling our agent's rational states.

The difficulty we have in deciding which impossible worlds may be accessible to our non-idealized agents is due to the fact that, seemingly, there is no third option: either epistemically accessible worlds are closed under full logical consequence, or they turn out to be obviously impossible. That, in short, is what Jago (2014a) calls *Bjerring's Problem* (which we'll discuss in more detail in §10.3).

(p.119) What seems especially difficult to do is to model epistemic agents that are rationally *competent*, in spite of not being omniscient. Real agents just cannot believe all that follows from what they believe. But if 'anything goes', so that believing something does not entail believing anything else in particular, then it seems that we are modelling agents who are not even moderately rationally competent: they fail to believe obvious consequences of what they believe.

One recent approach to the issue develops an idea for modelling competent but non-omniscient agents dynamically, in terms of how their beliefs will or may evolve over time due to epistemic actions and events. As a response to the problem of modelling competent but non-omniscient agents, the idea was originally put forward by Duc (1995, 1997). Bjerring and Rasmussen (2018) and Rasmussen (2015) update the approach, using the *dynamic epistemic logics* framework. To evaluate their approach, we'll need to understand a little about dynamic epistemic logic.

Dynamic logics contain operators based on actions. If a is an action, then '$[a]$' is an operator, and '$[a]A$' says that, after action a has been carried out, A is the case. Semantics for these operators is given in terms of transformations on the

model. Typically, the semantics of such logics uses pointed Kripke models, that is, models $\mathscr{M}^w$ with a regular Kripke model $\mathscr{M}$ and a 'designated' base world w, which is itself in $\mathscr{M}$ and can be thought of as the world considered to be the actual one. Then $[a]A$ is true in $\mathscr{M}^w$ iff A is true in all pointed models $\mathscr{N}^{w_1}$ obtained by transforming $\mathscr{M}^w$ according to the instructions encoded in action a. Dynamic epistemic logic, developed by Baltag et al. (1998), Segerberg (1995), Van Benthem (2011), Van Ditmarsch (2005), and Van Ditmarsch et al. (2008), adds epistemic operators to the dynamic semantics. Baltag and Renne (2016) give an introduction to the approach.

Bjerring and Rasmussen (2018), following ideas in Rasmussen 2015, adapt this approach to model competent but non-omniscient agents. In their models, agents count as competent insofar as they unfold the consequences of their beliefs, up to a certain 'depth' of reasoning. Their key dynamic operator is of the form '$\langle n\rangle A$', to be **(p.120)** read as: 'After n steps of logical reasoning, A may be the case'. These steps of reasoning are n applications of rules from a chosen set $\mathscr{R}$. The approach also has an epistemic operator B, allowing for sentences of the form '$\langle n\rangle BA$', saying that the agent can come to believe that A after n steps of reasoning.

Models (adapted to our own notation in this book) are tuples $\mathscr{M} = \langle\!\langle\, W, N, f, v \,\rangle\!\rangle$, where W is the total set of worlds, $N \subseteq W$ is the set of normal-possible worlds, f is an epistemic projection function (see §5.1) mapping each world to the set of worlds epistemically accessible from it, and v is a valuation function. Pointed models are model-world pairs, written $\mathscr{M}^w$, where $w \in N$.

Given a pointed model $\mathscr{M}^w$ and a set of rules $\mathscr{R}$, we can define a set of epistemic projection functions $\mathscr{F}^n$ for each integer n. Intuitively, these functions capture all chains of reasoning of length n using rules in $\mathscr{R}$, thus, all ways in which the agent can modify the set of worlds initially seen as epistemically possible, by performing a chain of reasoning steps of length n. We then define an equivalence relation $\sim^n$ for each n, relating pointed models that differ at most in their projection functions f, which must be chosen from $\mathscr{F}^n$. (We skip the definitions here: see Bjerring and Rasmussen (2018) for the details. We'll present a related rule-based approach in detail in §10.5.)

We then extend each pointed model $\mathscr{M}^w$'s valuation function v to all formulas. The clauses for connectives and B are as in standard epistemic logic, and the clause for $\langle n\rangle A$ is given as follows (at $w \in N$; impossible worlds have complex formulas evaluated directly, non-recursively):

($\mathrm{S}\langle n\rangle$) $v_w(\langle n\rangle A) = 1$ iff there is some pointed model $\mathscr{N}^w$ with extended valuation v' such that $\mathscr{M}^w \sim^n \mathscr{N}^w$ and $v'_w A = 1$.

The Bjerring-Rasmussen framework is a plausible and promising model of how non-omniscient agents can deductively unfold the consequences of their beliefs, and update their belief states accordingly. For, given a set of inference rules, we can align what the agent *can* come to believe in n steps of reasoning with possible proofs containing **(p.121)** no more than n steps. Bjerring and Rasmussen then prove that, if C follows from $A_1, \ldots, A_m$ in n steps of reasoning given rules $\mathscr{R}$, then $BA_1, \ldots, BA_m$ together entail $\langle n \rangle BC$ (Bjerring and Rasmussen 2018, 17, Corollary 1).

Their overall aim, however, is to capture a non-omniscient agent's logical competence, for which they provide the following behavioural test:

> For any p and q such that q follows trivially from p, if an agent believes p, then upon being asked whether q is the case does she immediately answer 'yes'? If she does, she passes the test and counts as moderately logically competent.
>
> (Bjerring and Rasmussen 2018, 3)

One might quibble with this test (after all, agents can be confused about what they believe; they often speak insincerely; and often require a moment's reflection before asserting). But let's accept it for our evaluation of the approach.

Does Bjerring and Rasmussen's approach pass the test they set? This is formulated in terms of what an agent will do: she will answer 'yes' when asked whether q, a trivial consequence of her beliefs. But the formal approach tells us about what an agent *can* come to believe, within n steps of reasoning. The main result is that $BA_1, \ldots, BA_m$ together entail $\langle n \rangle BC$ when C follows from the A_is within n steps. $\langle n \rangle BC$ says that the agent can come to believe C within n steps: there is some n-step chain of reasoning the agent can follow, via which she will come to believe that C. This isn't the same as telling us that our agent will come to believe that C. For she might follow some other chain of reasoning, and thereby come to believe something other than C. So there's no guarantee that the modelled agent will answer 'yes' when asked whether q. As such, that agent hasn't been show to pass the test for moderate logically competence.

If we're interested only in agents which pass Bjerring and Rasmussen's test, then we need to focus on what the modelled agent *will* come to believe after n steps, however she reasons. We need to switch from particular to universal quantification over pointed models, that **(p.122)** is, to a box-like dynamic modality, '$[n]$'. $[n]A$ is true on a pointed model when A is true on all $\sim^n$-related pointed models. (Then $[n]A$ is equivalent to $\neg\langle n \rangle \neg A$. But we can't define $[n]A$ directly in this way, since $[n]A$ and $\neg\langle n \rangle \neg A$ may come apart at impossible worlds.)

It's easy to see that, if the rules in $\mathscr{R}$ are sufficiently general, then there's nothing the modelled agent *must* believe after n steps (for any finite n). Suppose $\mathscr{R}$ contains Disjunction Introduction and our agent believes that p. Then one chain of reasoning goes: p, $p \vee p$, $p \vee p \vee p$, and so on. Another goes: p, $p \vee q$, $p \vee q \vee q$, and so on. The agent can go for either chain of reasoning. So the only thing she is guaranteed to believe after n steps of reasoning is p, which she believed to begin with. In other words, $[n]BA$ entails BA.

It also turns out that the move from one model to a $\sim^n$-related one is belief-monotonic, which means that an agent's current beliefs are preserved (with additions, but no subtractions) in the new model. This guarantees that BA entails $[n]BA$, and hence that $[n]BA$ is equivalent to BA. So formulas of the form '$[n]BA$' tell us nothing about logical competence, over and above what formulas of the form 'BA' tell us. But then, the modelled agent will fail the test for moderate logical competence, since beliefs aren't closed under any notion of consequence. Neither dynamic modality, then, helps us capture the target concept of moderate logical competence.

Bjerring and Rasmussen's behavioural test is (rightly, in our opinion) a normative one. It sets up a standard to be met, and so tells us something about what's expected of agents. Ordinary reasoners can fail the test, of course. We all make mistakes in our reasoning from time to time, and thereby, on those occasions, fail to live up to rational standards. In a normative setting, the aim is not to model what an agent can come to believe. Rather, it's to model a normative notion of belief, which builds in a certain amount of rationality, without idealizing agents to the point of logical omniscience.

Much more needs to be said on the key ideas we've introduced in this section. One is that of a *trivial inference*. We offer an analysis in §9.5. Another is the idea that epistemically accessible worlds cannot be obviously impossible. This leads to Bjerring's Problem: that **(p.123)** worlds are either closed under full logical consequence, or obviously impossible, and neither should be epistemically accessible. We offer a philosophical analysis of the problem, and attempt to draw a distinction between *obvious* and *subtle* impossibilities within a formal model, in §9.5. Finally, there is the normative, rational, but non-ideal notion of belief we've just discussed. We propose a model of the notion in §10.5, based on the idea of subtly impossible worlds from §9.5.

Chapter Summary

Standard possible-worlds epistemic logic gives rise to the problem of logical omniscience (§5.1). There are attempts to deal with the problem without using impossible worlds. We discussed a number of these approaches, and found them all wanting (§5.2). The impossible worlds approach is immediately more successful, but faces a deep problem: how should impossible worlds be constrained, so as to give adequate models of knowledge and belief (§5.3)? One

option is to take impossible worlds to be closed under some weaker-than-classical logic. But this approach does not genuinely solve the problem of logical omniscience (§5.4).

A different approach is the *dynamic* one, whereby epistemic states are not closed at any one time, but nevertheless evolve towards closure in a dynamic way (§5.5). We found this approach promising, but we will propose an alternative philosophical account of epistemic and doxastic states in §10.5. **(p.124)**