# A dynamic solution to the problem of logical omniscience

[Author name(s)]

[Date]

**Abstract**

The standard possible worlds model of belief is committed to logical omniscience, whereby agents are characterized as believing all logical consequences of what they believe, including all logical truths. Logical omniscience is a problem if we want to reason about the doxastic lives of agents who—much like ordinary human beings—are not logically omniscient, but nevertheless logically competent. Many proposed solutions to the problem of logical omniscience center around the use of impossible worlds, where the laws of classical logic fail to hold. In this paper, we argue first that standard versions of the impossible worlds model of belief fail to describe agents who are both logically non-omniscient and logically competent. In light of these negative arguments, we propose to *dynamize* the impossible worlds framework: it must allow us to model not only what agents believe, but also what they believe after some chain of logical reasoning. We then develop a dynamic impossible worlds model of belief and show that it successfully characterizes agents as both logically non-omniscient and logically competent.

**Keywords** Logical omniscience · Doxastic logic · Epistemic logic · Impossible worlds · Resource-bounded reasoning · Bounded rationality

# 1 Introduction

Consider the standard possible worlds semantics for belief, which dates back to Hintikka's *Knowledge and Belief* (1962):

> (**Belief**) An agent believes a proposition $p$ iff $p$ is true at all possible worlds that are doxastically possible for the agent.

Since we normally take possible worlds to be *logically* possible, it follows from (Belief) that agents are *logically omniscient*: they believe all logical consequences of what they believe, including all classical tautologies.[1] For suppose that an agent believes $p$ and consider any $q$ that is logically entailed by $p$. By (Belief), $p$ is true at all doxastically possible worlds for the agent. Since $p$ entails $q$, all logically possible worlds that verify $p$ also verify $q$. Given the assumption that all doxastically possible worlds are also logically possible, it follows that $q$ is true at all doxastically possible worlds for the agent. Hence, by (Belief), the agent believes $q$. So if an agent believes $p$, she believes any logical consequence of $p$.

Logical omniscience is a problem if we are interested in the doxastic lives of agents who—much like ordinary human beings—fail to believe all the logical consequences of what they believe. It may well be that Goldbach's Conjecture follows logically from the Peano axioms. Yet, just because you believe that the Peano Axioms are true, you need not believe that Goldbach's Conjecture is true. Similarly, logical omniscience is a problem if we are interested in artificially intelligent agents—such as computers and robots—that are subject to time constraints and limited computational resources. While such agents can compute *some* logical consequences of what is stored in their memory, they cannot generally compute *all* such consequences.

Also, logical omniscience is a problem if we are interested in a normative theory of belief that is sensitive to an agent's limited capacities for inferential reasoning. While computationally limited agents fail to live up to the

[1] Here we use the standard definition of logical omniscience according to which beliefs are closed under logical consequence in classical logic. Several weaker notions of logical omniscience have been discussed in the literature, but our solution will straightforwardly apply to those as well. For an overview of standard approaches to logical omniscience, see Fagin et al. (1995, ch. 9).

high rationality standards for logically omniscient agents, they need not, as Weirich points out, "be irrational in any way. They may fully conform to all standards for agents with their limitations" (Weirich 2004, p. 100). For the purposes of developing a theory of boundedly rational belief, logical omniscience is inadmissible.

If our sole goal is to describe agents who fall short of logical omniscience, our job is done once we have a model of agents who fail to believe all logical consequences of what they believe. As we shall see in section 2, there are worked out formal theories of belief that meet this objective. But if, like Chalmers (2011), Cherniak (1986), Jago (2013; 2014), Weirich (2004) and others, our goal is to describe agents who, despite being logically non-omniscient, are nevertheless logically *competent*, there is more work to be done. For, as we will shall see, it is a non-trivial task to avoid logical omniscience without sacrificing logical competence.

When is an agent logically competent? As a rough first pass, we can say that an agent is logically competent when she does not "miss out" on any *trivial* logical consequences of what she believes. To give us a handle on this intuitive characterization, we can appeal to the following behavioral 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 logically competent. For instance, suppose you believe that it rains and that if it rains, the streets are wet. We can then ask: will you answer "yes" when asked whether the streets are wet? Assuming that you are sincere, attentive, mentally well-functioning, and so on, it surely seems so. So you do not miss out on this trivial logical consequence of your beliefs, and thus count as logically competent.

Many of the reasons for being interested in logically non-omniscient agents are also reasons for being interested in agents who do not miss out on any trivial logical consequences of their beliefs. For instance, suppose we aim to model agents such as humans and computers who generally have an ability to engage in at least simple or trivial logical reasoning. You are about to leave your house and notice that it snows heavily outside. Well aware of your fragile health, you believe that if it snows heavily, you should wear a winter jacket when outside. Surely, in the normal run of things, you will wear the winter jacket when leaving the house. To explain this behavior, we must appeal not only to your desire not to get sick, but also to your capacity for simple logical reasoning. For based on your beliefs that it snows heavily and that if it snows heavily, you should wear a winter jacket, you can infer the trivial consequence that you should wear a winter jacket. Had you missed out on this trivial consequence of your beliefs, you would not have been able to act in ways that would satisfy your desire not to get sick. To be sure, our claim is not that humans never make mistakes even in simple logical reasoning. But we take it that most people in ordinary circumstances have at least a basic ability—albeit a fallible one—to engage in simple or trivial logical reasoning. And it is this ability we want to capture by developing a model of agents who do not miss out on any trivial consequences of what they believe.

Also, by developing a model of agents who do not miss out on any trivial consequences of their beliefs, we can give an account of rational belief that is sensitive to the limited resources that agents such as humans and artificial agents have available for logical reasoning. Following Cherniak (1986) and others, we might plausibly deny that *any* set of propositions can characterize the doxastic state of a boundedly rational agent, even if we give up the ideal that only logically consistent and closed such sets can. For instance, we might hold that a boundedly rational agent must be able to infer at least the *trivial* consequences of her beliefs, and be able to eliminate at least the *trivial* inconsistencies from her doxastic state. To capture such constraints on non-ideal rationality, avoiding logical omniscience is not enough: we also need the tools to model agents who have the ability to tease out at least the simple or trivial logical consequences of their beliefs.

What counts as a "trivial" logical consequence depends on the cognitive or computational resources that agents have available for logical reasoning. If you are an experienced logician, it might be trivial for you to see that $\neg q \rightarrow \neg p$ entails $\neg(p \wedge \neg q)$, whereas, for a first-year philosophy student, this inference may be highly non-trivial. To capture this, we will adopt a simple "step-based" picture of what it means to reason with limited cognitive or computational resources.[2] On this picture, agents reason logically by applying rules from a set $\mathcal{R}$ of inference rules, and one "step" of logical reasoning corresponds to one application of a rule in $\mathcal{R}$. This allows us to model an agent's computational resources in terms of the number $n$ of steps of reasoning that the agent can trivially perform using the rules in $\mathcal{R}$. In the limit where $n = 0$, agents have no available computational resources, and thus no logical consequence counts as trivial (assuming that nothing is provable in zero steps of reasoning). In the opposite limit where $n$ approaches infinity, agents have unlimited computational resources, and thus every logical consequence, however complicated, counts as trivial. In between these extremes, we find agents with an intermediate amount of resources available for trivial logical reasoning. For such agents, some but not all logical consequences count as trivial.

We can then define trivial logical consequence in terms of $n$-step inferability in $\mathcal{R}$:

> **(Trivial consequence)** For any propositions $p$ and $q$, $q$ is a trivial logical consequence of $p$ iff $q$ can be inferred from $p$ by at most $n$ applications of the rules in $\mathcal{R}$.

When $q$ is a trivial logical consequence of $p$, we will also say that a chain of logical *reasoning* from $p$ to $q$ is trivial. We can then stipulatively understand 'logically competent agents' as agents who have the ability to infer at least the trivial logical consequences of what they believe. Since such agent have the ability to tease out what follows within $n$ steps of logical reasoning from what they believe, they do not—in any relevant sense—miss out on any trivial consequences of what they believe.

According to (Trivial consequence), what counts as a trivial consequence depends not only on the value of $n$, but also on the rules in $\mathcal{R}$. For instance, if $\mathcal{R}$ contains only modus tollens, then $\neg p$ is the only trivial consequence of $\{p \rightarrow q, \neg q\}$, no matter how high $n$ goes. If, on the other hand, $\mathcal{R}$ corresponds to, say, a complete proof system for classical propositional logic such as

[2] This step-based approach is inspired by so-called *active logics* (formerly known as *step logics*); see Drapkin et al. (1999) and Drapkin & Perlis (1986; 1990).

natural deduction, the number of trivial consequences of $\{p \to q, \neg q\}$ will be significantly larger, for sufficiently high values of $n$.

We will deliberately leave the content of $\mathcal{R}$ and the value of $n$ unspecified. By treating $\mathcal{R}$ and $n$ as unspecified variables, the framework below can be applied in a wide range of theoretical contexts that call for different specifications of $\mathcal{R}$ and $n$. For instance, if we are interested in modelling rather complex agents with a powerful reasoning mechanism and high computational resources, we can let $\mathcal{R}$ constitute a complete proof system for propositional logic and let the value of $n$ be high. If, on the other hand, we are interested in modelling relatively simple agents with a potentially very weak reasoning mechanism—such as those modeled by Konolige (1986)—and only few computational resources, we can give a corresponding sparse characterization of $\mathcal{R}$ and low value of $n$.

Our aim in this paper, then, is to develop a formal model of logically non-omniscient, yet logically competent agents. We will build our model on the traditional impossible worlds model of belief developed by Cresswell (1970; 1972; 1973), Fagin et al. (1995), Hintikka (1975), Rantala (1982), Wansing (1990), and others. Subtleties aside, such models retain (Belief) as the semantics for belief but extend the underlying space of worlds to include *impossible worlds*, where the laws of classical logic fail to hold—we return to the details in section 2.

We proceed as follows. In section 2, we argue that standard versions of the impossible worlds model of belief fail to describe agents who are both logically non-omniscient and logically competent. In light of these negative arguments, we argue in section 3 that we need to *dynamize* the impossible worlds framework: it must allow us to model not only what agents *believe*, but also what they believe after a given (trivial) chain of logical *reasoning*. We develop a dynamic impossible worlds model of belief and show that it successfully describes agents who are both logically non-omniscient and logically competent. In section 4, we discuss Mark Jago's (2013; 2014) recent impossible worlds model of belief that, like ours, promises to steer clear of both logical omniscience and logical incompetence. We argue that our model is superior to his. Finally, in section 5, we conclude.

## 2 Impossible worlds and logical omniscience

The impossible worlds model of belief retains (Belief) but extends the underlying space of worlds to include what Hintikka called *impossible possible worlds*: worlds that "look possible and hence must be admissible as epistemic alternatives but which none the less are not logically possible" (Hintikka 1975, p. 477). The motivating thought is that, for agents with limited computational resources, the space of doxastic possibilities is larger than the space of logical possibilities. Such agents may well believe each of the Peano Axioms but fail to believe Goldbach's Conjecture, in which case the Peano Axioms are true in all doxastic alternatives for them, whereas Goldbach's Conjecture is not. Since no logically possible world verifies the Peano Axioms but falsifies Goldbach's Conjecture—assuming that the Conjecture is indeed entailed by the Peano Axioms—it follows that doxastic possibility outstrips logical possibility. By including logically impossible worlds that verify the Peano Axioms but not Goldbach's Conjecture, we can model agents for whom both the Conjecture and its negation are live doxastic possibilities.

Consider then the standard impossible worlds semantics for belief:

> (**Belief-impossible**) An agent believes a proposition $p$ iff $p$ is true at all worlds (whether possible or impossible) that are doxastically possible for the agent.

Note that (Belief-impossible) says nothing about the nature of impossible worlds. Corresponding to different conceptions of what these worlds are, we get different versions of the impossible worlds approach. On one conception—what Berto (2013) calls the "American stance"—impossible worlds are not closed under *any* notion of logical consequence: for *any* set of propositions, however blatantly inconsistent or logically ill-behaved, some impossible world verifies just those propositions. On a second conception—what Berto (2013) calls the "Australasian stance"—impossible worlds are closed under logical consequence in some non-classical logic: they verify only sets of propositions that respect the laws of the non-classical logic. Yet, as we shall argue now, neither conception allows us to model agents who are both logically non-omniscient and logically competent.

*The American stance.* Suppose first that impossible worlds are not closed under any notion of logical consequence. Impossible worlds then satisfy the following principle:

> **(Non-closure)** For any propositions $p$ and $q$, some impossible world verifies $p$ but not $q$.

If we adopt (Non-closure), we avoid logical omniscience. For given that some impossible world verifies $p$ but not $q$, for any $p$ and $q$, it need not follow that all doxastically possible worlds that verify $p$ also verify $q$. In turn, by (Belief-impossible), agents can believe $p$ but fail to believe $q$. So agents need not believe all logical consequences of what they believe.

While (Non-closure) can help us avoid logical omniscience, it does not give us the tools to model logically competent agents. Since (Non-closure) implies that agents can believe $p$, but fail to believe $q$, for any $p$ and $q$, agents need never believe *any* logical consequences of what they believe, nor are they described as having the ability to form such beliefs. So nothing in the formalism predicts that agents will ever pass the test for logical competence: if asked about some trivial consequence of their beliefs, they need not assent since they might fail to believe the trivial consequence in question.[3]

*The Australasian stance*: Suppose next that impossible worlds are closed under logical consequence in some non-classical logic $L$ (e.g. intuitionistic or paraconsistent logic).[4] Impossible worlds then satisfy the following principle:

[3]Note that nothing in (Non-closure) prevents us from specifying particular models of agents who believe some logical consequences of what they believe by "writing in by hand" which worlds are accessible from which. Yet, as we will argue in section 2, it is not possible, even by hand, to specify models of logically non-omniscient agents who believe *all trivial* logical consequences of what they believe. As such, we cannot construct by hand impossible worlds models that successfully describe logically non-omniscient, but logically competent agents. Also, in a broader picture, the strategy of handpicking individual models to fit one's particular purposes does not square well with the usual way in which model-theoretic tools are used to capture general properties of modal concepts. For instance, if we want knowledge to satisfy the KK-principle, we usually require that *all* models have a transitive accessibility relation. Without this general constraint, we could not prove that $Kp$ entails $KKp$. In the same way, without any general constraints on models, we cannot establish formal results about the beliefs of logically competent agents. Thanks to an anonymous reviewer for raising these issues.

[4]See Fagin et al. (1995), Levesque (1984), and Lakemeyer (1987) for approaches to logical omniscience that appeal to a non-classical logic.

**(Non-classical closure)** For any propositions $p$ and $q$ such that $q$ is a logical consequence of $p$ in $L$, any impossible world that verifies $p$ verifies $q$.

Given that $L$ is weaker than classical logic, (Non-classical closure) avoids describing agents as logically omniscient with respect to *classical* logic. For consider some $p$ and $q$ such that $q$ follows from $p$ in classical logic, but not in $L$. By (Non-classical closure), impossible worlds may then verify $p$ but fail to verify $q$. In turn, agents may believe $p$ but fail to believe $q$. So agents need not believe all *classical* consequences of what they believe.

However, agents are still characterized as logically omniscient with respect to the non-classical logic $L$. For instance, if we understand $L$ as a paraconsistent logic, agents will believe all paraconsistent consequences of what they believe, including all paraconsistent tautologies. But just as it is implausible that computationally limited agents believe all *classical* consequences of what they believe, it is implausible that they believe all *paraconsistent* consequences of what they believe. For even supposing that such agents reason using paraconsistent logic, they cannot—due to their limited resources—reason *unlimited* in that logic. So (Non-classical closure) still commits us to an undesirable kind of logical omniscience.

Moreover, (Non-classical closure) does not allow us to capture adequately the sense in which agents are logically competent. For not all agents have a non-classical reasoning mechanism, let alone the *same* reasoning mechanism. Consider for instance a proof generator for classical propositional logic. While such a generator reasons purely classically, it nevertheless falls short of logical omniscience due to its limited computational resources. Describing that agent as omniscient with respect to some non-classical logic misses the target: the agent is neither logically omniscient with respect to any logic, nor does the agent reason non-classically.

In light of the problems with (Non-closure) and (Non-classical closure), it is natural to consider a closure principle on impossible worlds that trades on our definition of trivial logical consequence:

**(Trivial closure)** For any $p$ and $q$ such that $q$ is a trivial logical consequence of $p$ (i.e. such that $q$ is $n$-step inferable in $\mathcal{R}$ from $p$), any impossible world that verifies $p$ also verifies $q$.

By closing impossible worlds under trivial logical consequence, we ensure that agents believe all trivial consequences of what they believe. For consider an agent who believes $p$, and let $q$ be any trivial logical consequence of $p$. By (Belief-impossible), $p$ is true at all worlds that are doxastically possible for the agent. By (Trivial closure), $q$ must then also be true at all such worlds. So, by (Belief-impossible), the agent believes $q$.

Needless to say, if we embrace (Trivial closure), we must still ensure that impossible worlds are not fully deductively closed with respect to $\mathcal{R}$. For if they are, we end up describing agents as logically omniscient with respect to $\mathcal{R}$. So the challenge is to satisfy both (Trivial closure) and the following principle:

> **(Deductive openness)** For some impossible world $w$ and proposition $p$, the set of true propositions at $w$ entails $p$ in some number of steps in $\mathcal{R}$, but $w$ does not verify $p$.

As it turns out, however, it is impossible to satisfy both (Trivial closure) and (Deductive openness).[5] To see this, consider any inference in $\mathcal{R}$ from a set $\Gamma$ of premises to a conclusion $q$. In order for (Trivial closure) and (Deductive openness) to be jointly satisfiable, there must be at least *one* impossible world $w$ such that:

(i) $w$ verifies all the premises in $\Gamma$,
(ii) $w$ verifies every proposition that is a trivial logical consequence of the set of true propositions at $w$, and
(iii) $w$ does not verify the conclusion $q$.

But we can now show that no world can jointly satisfy (i) to (iii). Note first that since $\Gamma$ entails $q$ in $\mathcal{R}$, there exists a sequence of propositions $\langle \Gamma, q_1, q_2, \ldots, q \rangle$ corresponding to an inference from $\Gamma$ to $q$ by some number

[5] Earlier versions of this argument can be found in [Author] (2013; 2014), [Author] (2016), and Jago (2014).

of applications of the rules in $\mathcal{R}$.[6] Given that $n \geq 1$—i.e. given that agents meet a minimal threshold of logical competence—it follows by (Trivial consequence) that each $q_i$ is a trivial consequence of $q_{i-1}$. Consider, then, a world $w$ that satisfies (i) and (ii): $w$ verifies each premise in $\Gamma$ as well as every trivial logical consequence of the set of true propositions at $w$. It then follows that $w$ verifies $q_1$. If it did not, it would fail to verify a trivial consequence of $\Gamma$ and hence fail to satisfy (ii). But given that $w$ verifies $q_1$, it must also verify $q_2$ since $q_2$ is a trivial logical consequence of $q_1$. Continuing this line of reasoning, it follows that $w$ must verify $q$. So if a world satisfies (i) and (ii), it cannot satisfy (iii). So (Trivial closure) and (Deductive openness) cannot be jointly satisfied: if we close worlds under trivial logical consequence in $\mathcal{R}$, we end up closing them under full logical consequence in $\mathcal{R}$. Intuitively, a world that is closed under trivial logical consequence *collapses* under its own deductive weight to a world that is closed under full logical consequence.

More generally, as Jago (2014) has also pointed out, this "collapse result" can be established without appeal to any particular formal theory of belief. For given the line of reasoning above, it is straightforward to see that no agent can satisfy the following conditions:

(1) The agent believes all the premises in $\Gamma$,
(2) the agent believes all trivial consequences of what she believes, and
(3) the agent does not believe the conclusion $q$.

When both (1) and (2) are satisfied, (3) cannot be. That is, if an agent believes every proposition that follows trivially from her beliefs, she ends up believing *all* consequences of what she believes. Thus we cannot model agents who are both logically non-omniscient and logically competent by closing beliefs under trivial logical consequence.

[6] While the details of the inference from $\Gamma$ to $q$ depends on the rules in $\mathcal{R}$, nothing of importance hinges on whether we think of $\mathcal{R}$ as a natural deduction system, a sequent calculus, or some other standard proof system. For as shown below, to establish that (i)-(iii) are not jointly satisfiable, we only need the assumption that there exists an inference in $\mathcal{R}$ from $\Gamma$ to $q$ such that each step in the inference is trivial. And at least for standard rules such as conjunction introduction, modus ponens, and double negation elimination, it is plausible that each such rule is cognitively or computationally trivial to apply. For further motivation of these thoughts, see also Author (2013) and Jago (2014).

How then should we model agents who are both logically non-omniscient and logically competent? We give our answer in the next section.

# 3 A dynamic model of belief

Consider a logically non-omniscient agent who passes the behavioral test for logical competence: for any $p$ and $q$ such that $q$ is a trivial consequence of $p$, if the agent believes $p$, then upon being asked whether $q$ is the case, she answers "yes" immediately. In light of the collapse result, we cannot explain the agent's behavior by assuming that she believes $q$ *prior* to being asked about it. For if she did, we know that she would believe all logical consequences of her beliefs. Instead we can explain how she passes the test by citing her ability to engage in trivial logical reasoning. When asked about $q$, the question primes her to perform the trivial inference from $p$ to $q$. This reasoning issues an immediate *transition* from a belief state that contains $p$ to one that also contains $q$. And when asked about $q$, the agent bases her positive answer on the updated belief state that contains $q$.

Hence we suggest that the proper way to model agents who are both logically non-omniscient and logically competent should appeal to a *relation* between distinct doxastic states. We will understand this relation dynamically as a *reasoning process* that issues a transition from a doxastic state containing the premises of a given inference to an updated doxastic state that also contains the conclusion.[7] In this way, the formalism below will enable us to model how an agent's beliefs can change as a result of exercising her ability to engage in trivial logical reasoning.

To formalize these ideas, we will develop a *dynamic* impossible worlds model of belief that describes not only what an agent believes, but also what she believes after having performed a trivial chain of logical reasoning. We will approach matters from a model-theoretic point of view and postpone

[7] The idea of modeling belief change in terms of transitions or relations between doxastic states is well-known from dynamic epistemic logic; see, for instance, van Ditmarsch et al. (2008) and Duc (1997).

proof-theoretic details for future work.[8]

First, recall that a chain of logical reasoning from $p$ to $q$ counts as trivial just in case $q$ can be inferred from $p$ by at most $n$ steps of logical reasoning using the rules in $\mathcal{R}$. Formally, we write '$\Gamma \vdash^n_{\mathcal{R}} \Gamma'$' to mean that $\Gamma'$ is $n$-step inferable from $\Gamma$ using the rules in $\mathcal{R}$, where $\Gamma$ and $\Gamma'$ are sets of sentences. For the central results below, we assume that $\vdash^n_{\mathcal{R}}$ is monotonic:

**($\mathcal{R}$-monotonicity)** If $\Gamma \subseteq \Gamma'$ and $\Gamma \vdash^n_{\mathcal{R}} p$, then $\Gamma' \vdash^n_{\mathcal{R}} p$.

($\mathcal{R}$-monotonicity) ensures that inferences are never defeated by adding extra premises.

We define our logic over a propositional modal language $\mathcal{L}$:

**Definition 1.** ***(Language)*** *The language $\mathcal{L}$ is defined inductively from a set $\Phi$ of atomic sentences, an adequate set $\{\neg, \wedge\}$ of boolean connectives (from which $\vee, \rightarrow$, and $\leftrightarrow$ are defined in the usual way), a doxastic operator $B$, and a countably infinite set of dynamic operators $\langle n \rangle$ and $[n]$:*

$$p \ ::= \ \varphi \mid \neg p \mid p \wedge q \mid Bp \mid \langle n \rangle p \mid [n]p,$$

*where $n = 0, 1, 2, \ldots$ and $\varphi \in \Phi$.*

The operators in $\mathcal{L}$ have the following intuitive readings:

$Bp$: The agent believes $p$.
$\langle n \rangle p$: After some $n$ steps of logical reasoning, $p$ is the case.
$[n]p$: After any $n$ steps of logical reasoning, $p$ is the case.

In particular, $\langle n \rangle Bp$ says that 'the agent believes $p$ after some $n$ steps of logical reasoning' or, equivalently, that 'the agent believes $p$ after some trivial chain of logical reasoning.' Similarly, $[n]Bp$ says that 'the agent believes $p$ after any $n$ steps of logical reasoning' or, equivalently, that 'the agent believes $p$ after any trivial chain of logical reasoning.'

We define doxastic models as follows:

[8]For some preliminary work in this direction, see Duc (1997) and [Author] (2015).

**Definition 2.** ***(Doxastic model)*** *Let $W^P$ and $W^I$ be non-empty sets of possible and impossible worlds respectively, and let $W = W^P \cup W^I$. A doxastic model for a single agent is a structure:*

$$M = \langle W^P, W^I, f, V\rangle,$$

*where $f : W \mapsto 2^W$ is an accessibility function, assigning to each world in $W$ a set of worlds in $W$, and $V : W \mapsto 2^{\mathcal{L}}$ is a function, assigning to each world in $W$ a set of sentences in $\mathcal{L}$.*

Here $f(w)$ corresponds to the set of doxastically accessible worlds from $w$, and $V(w)$ corresponds roughly to the set of sentences that are true at $w$.[9] Whereas possible worlds are complete and logically consistent entities, impossible worlds need neither be complete nor subject to any closure conditions. However, we understand impossible worlds in a way that deviates slightly from (Non-closure) since we require that they are minimally consistent:

> **(Minimal consistency)** For any world $w \in W^I$ and sentence $p \in \mathcal{L}$, $\{p, \neg p\} \notin V(w)$.

By excluding worlds from $W$ that verify explicit contradictions of the form $\{p, \neg p\}$, we ensure that agents never believe both $p$ and $\neg p$, for any $p$. While a complete model of logically non-omniscient agents should arguably explain how such agents can reason rationally in the presence of contradictions, a detailed analysis of the kinds of rules and principles that should govern contradiction-tolerant rational reasoning is well beyond the scope of this paper.

We can now go on to develop the semantics for $\mathcal{L}$. Since we retain (Belief-impossible) as our semantics for $Bp$, we need only to provide a semantics for the new dynamic operators $\langle n\rangle p$ and $[n]p$. To flag the core idea behind the semantics for $\langle n\rangle p$, focus, as a special case, on the sentence type $\langle n\rangle Bp$. The

[9] Less roughly, as we shall see below, for $w \in W^P$, $V(w)$ yields the set of atomic sentences that are true at $w$, while the standard recursive satisfaction clauses in turn yield the truth values of all complex sentences at $w$. For $w \in W^I$, however, $V(w)$ yields the set of *all* true sentences at $w$, whether atomic or complex.

semantics below will tell us that $\langle n \rangle Bp$ is true at a world $w$ just in case $p$ follows within $n$ steps of logical reasoning from each doxastically accessible world from $w$. The truth conditions for $\langle n \rangle Bp$ will thus be weaker than those for $Bp$: while the semantics for $Bp$ requires that $p$ is *true* at all doxastically accessible worlds, the semantics for $\langle n \rangle Bp$ will merely require that *$p$ follows* from each doxastically accessible world within $n$ steps of logical reasoning in $\mathcal{R}$.

So, to specify the semantics, we first need a device that tells us what follows within $n$ steps of reasoning from a world:

**Definition 3.** ***($n$-radius)*** *The $n$-radius $w^n$ of a world $w \in W$ is defined as:*

$$w^n = \{w' | V(w) \vdash^n_{\mathcal{R}} V(w')\}.$$

*Each member of $w^n$ is called an $n$-expansion of $w$.*

The $n$-radius of a world $w$ is the set of $n$-expansions of $w$, where $w'$ is an $n$-expansion of $w$ just in case $V(w')$ is $n$-step inferable from $V(w)$ using the rules in $\mathcal{R}$. Since possible worlds are deductively closed, $w^n = \{w\}$ for any $w \in W^P$. By contrast, for $w \in W^I$, the $n$-radius of $w$ may contain many different $n$-expansions of $w$.

For the semantics for $\langle n \rangle Bp$, we want to require that at least *one* $n$-expansion of each doxastically accessible world from $w$ verifies $p$. So we need a device that can pick out exactly one $n$-expansion of each world that is doxastically accessible from $w$:

**Definition 4.** ***(Choice function)*** *Let $\mathcal{C} : 2^{2^W} \mapsto 2^{2^W}$ be a choice function that takes a set $\mathcal{W} = \{W_1, \dots, W_n\}$ of sets of worlds as input and returns the set $\mathcal{C}(\mathcal{W})$ of sets of worlds which results from all the ways in which exactly one element can be picked from each $W_i \in \mathcal{W}$. Each member of $\mathcal{C}(\mathcal{W})$ is called a choice of $\mathcal{W}$.*

To illustrate, let $\mathcal{W} = \{\{w_1, w_2\}, \{w_3\}\}$. Since each choice of $\mathcal{W}$ corresponds to one way of picking out exactly one world from each member of $\mathcal{W}$, there are two choices of $\mathcal{W}$: we can either pick $\{w_1, w_3\}$ or $\{w_2, w_3\}$. So $\mathcal{C}(\mathcal{W}) = \{\{w_1, w_3\}, \{w_2, w_3\}\}$.

We can now use ($n$-radius) and (Choice function) to define a relation $\overset{n}{\sim}$ between pointed models, where a pointed model consists of a model and a world. When the relation $\overset{n}{\sim}$ holds between two pointed models $(M, w)$ and $(M', w')$, we write '$(M, w) \overset{n}{\sim} (M', w')$' and say that $(M', w')$ is *$n$-accessible* from $(M, w)$. Informally, if $(M, w)$ characterizes an agent's current belief state, we want to say that $(M', w')$ is $n$-accessible from $(M, w)$ just in case $(M', w')$ characterizes a belief state that the agent can enter from $(M, w)$ by performing $n$ steps of logical reasoning. To capture this idea, $(M', w')$ should be $n$-accessible from $(M, w)$ just in case the set of doxastically accessible worlds from $w$ in $M$ is replaced in $M'$ by a choice of $n$-expansions of the accessible worlds from $w$ in $M$. So we need a device that can replace a set of worlds with a choice of $n$-expansions of those worlds:

**Definition 5.** ***($n$-variation)*** *Let $M = \langle W^P, W^I, f, V \rangle$ be a model. $\mathcal{F}^n$ (for $n = 0, 1, 2, \dots$) is a function from pointed models to sets of accessibility functions defined as:*

$$\mathcal{F}^n(M, w) = \left\{ g | g(v) = \begin{cases} c & \text{for } v = w \\ f(v) & \text{for } v \neq w \end{cases} \right\},$$

*where $c \in \mathcal{C}(\{w'^n | w' \in f(w)\})$. Each member of $\mathcal{F}^n$ is called an $n$-variation of $f$.*

So an accessibility function $f'$ is an $n$-variation of $f$ just in case $f'(w)$ is a choice of $n$-expansions of $f(w)$. To illustrate, suppose $f(w) = \{\alpha_1, \alpha_2\}$, $\varepsilon_1 \in \alpha_1^n$, and $\varepsilon_2 \in \alpha_2^n$. If $f'(w) = \{\varepsilon_1, \varepsilon_2\}$, then $f'$ is an $n$-variation of $f$.

We can then define the $n$-accessibility relation as follows:

**Definition 6.** ***($n$-accessibility)*** *Let $M = \langle W^P, W^I, f, V \rangle$ and $M' = \langle W^{P'}, W^{I'}, f', V' \rangle$ be any two models. Then $(M, w) \overset{n}{\sim} (M', w')$ iff $w' = w$, $W' = W$, $V' = V$, and $f' \in \mathcal{F}^n(M, w)$.*

According to ($n$-accessibility), $(M', w')$ is $n$-accessible from $(M, w)$ just in case the set of doxastically accessible worlds from $w$ in $M$ is replaced in $M'$ by a choice of $n$-expansions of the accessible worlds from $w$ in $M$ (see figure

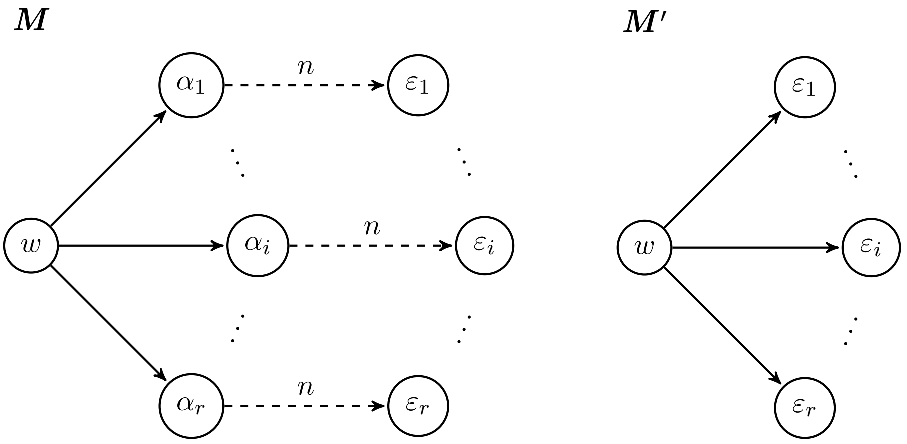


Figure 1: Illustration of ($n$-accessibility). A solid arrow from $w$ to $w'$ represents that $w'$ is doxastically accessible from $w$, and a dashed arrow labelled '$n$' from $w$ to $w'$ represents that $w'$ is an $n$-expansion of $w$. $(M', w)$ is $n$-accessible from $(M, w)$, since the set $\{\alpha_1, \ldots, \alpha_r\}$ of accessible worlds from $w$ in $M$ is replaced in $M'$ by a choice $\{\varepsilon_1, \ldots, \varepsilon_r\}$ of $n$-expansions of these accessible worlds.

1 for an illustration). We can think of the set of $n$-accessible pointed models as representing all the different ways in which an agent's doxastic state can change as a result of performing $n$ steps of logical reasoning.

Given ($n$-accessibility), we can now complete our semantics:

**Definition 7.** ***(Satisfaction)*** *Sentences are interpreted on pointed models* $(M, w)$*. We write* $M, w \vDash p$ *to mean that* $p$ *is true at* $w$ *in* $M$ *(or, equivalently, that* $(M, w)$ *satisfies* $p$*), and* $M, w \dashv p$ *to mean that* $p$ *is false at* $w$ *in* $M$*.*

*For* $w \in W^P$*:*

*(P1)* $M, w \vDash \varphi$ *iff* $\varphi \in V(w)$*, where* $\varphi \in \Phi$*.*
*(P2)* $M, w \vDash \neg p$ *iff* $M, w \nvDash p$*.*
*(P3)* $M, w \vDash p \wedge q$ *iff* $M, w \vDash p$ *and* $M, w \vDash q$*.*
*(P4)* $M, w \vDash Bp$ *iff* $M, w' \vDash p$ *for all* $w' \in f(w)$*.*
*(P5)* $M, w \vDash \langle n \rangle p$ *iff* $M', w' \vDash p$ *for some* $(M', w') : (M, w) \overset{n}{\sim} (M', w')$*.*
*(P6)* $M, w \vDash [n]p$ *iff* $M', w' \vDash p$ *for all* $(M', w') : (M, w) \overset{n}{\sim} (M', w')$*.*
*(P7)* $M, w \dashv p$ *iff* $M, w \nvDash p$*.*

*For* $w \in W^I$*:*

*(I1)* $M, w \vDash p$ *iff* $p \in V(w)$*.*

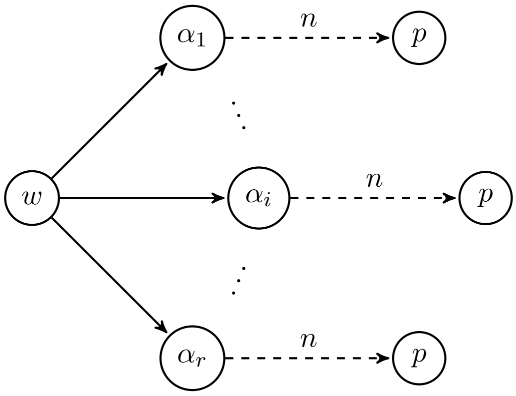


Figure 2: Illustration of the semantics for $\langle n \rangle Bp$. $\langle n \rangle Bp$ is true at $w$ since $p$ follows within $n$ steps of reasoning from each accessible world from $w$. Here the $p$s in the $n$-expansions of the accessible worlds $\{\alpha_1, ..., \alpha_r\}$ from $w$ are not names of worlds, but indicate that $p$ is true at the relevant $n$-expansion.

*(I2)* $M, w \dashv p$ *iff* $\neg p \in V(w)$.

For both possible and impossible worlds, $p$ is false at $w$ just in case $\neg p$ is true at $w$. At impossible worlds, however, $p$ may be neither true nor false, whereas possible worlds never contain truth value gaps. Validity is defined with respect to possible worlds only: $p$ is valid iff $p$ is true at all possible worlds in all models (across some class of models). (P4) is a formalization of (Belief-impossible) since the accessibility function $f$ ranges over both possible and impossible worlds. (P5) says that $\langle n \rangle p$ is true at $w$ in $M$ just in case $p$ is satisfied by at least one $n$-accessible pointed model from $(M, w)$. In particular, $\langle n \rangle Bp$ is satisfied by $(M, w)$ just in case $Bp$ is satisfied by some $n$-accessible pointed model from $(M, w)$ (see figure 2 for an illustration). Hence (P5) captures the central idea that an agent believes $p$ after some trivial chain of logical reasoning whenever there is a transition from the agent's doxastic state through $n$ applications of the rules in $\mathcal{R}$ to a state in which the agent believes $p$.[10] By contrast, (P6) gives the conditions under which an agent believes $p$ after *any* trivial chain of logical reasoning.

We can now establish the main result of the paper (see figure 3 for a

[10]So understood, the semantics for $\langle n \rangle p$ is structurally similar to other well-known semantics for dynamic operators that appeal to relations between or updates on epistemic models; see for instance the dynamic semantics for public announcement in van Ditmarsch et al. (2008, ch. 4).

diagrammatic representation):

**Theorem 1.** *If* $\{p_1, \ldots, p_k\} \vdash^n_{\mathcal{R}} q$ *and* $\langle m_i \rangle Bp_i$*, for* $1 \leq i \leq k$*, then* $\langle \omega + n \rangle Bq$*, where* $\omega = m_1 + \cdots + m_k$*.*

*Proof.* Let $M = \langle W^P, W^I, f, V \rangle$ be any model, and suppose $\{p_1, \ldots, p_k\} \vdash^n_{\mathcal{R}} q$ and $M, w \vDash \langle m_i \rangle Bp_i$, for $1 \leq i \leq k$. We must show that $M, w \vDash \langle \omega + n \rangle Bq$, where $\omega = m_1 + \cdots + m_k$. By (P5), $M_i, w_i \vDash Bp_i$ for some $(M_i, w_i) : (M, w) \overset{m_i}{\sim} (M_i, w_i)$, for $1 \leq i \leq k$. By ($n$-accessibility), $M_i, w \vDash Bp_i$ for some $M_i = \langle W^P, W^I, f_i, V \rangle$, where $f_i \in \mathcal{F}^{m_i}(M, w)$. By ($n$-variation), $f_i(w) = c_i$ for some choice $c_i \in \mathcal{C}(\{v^{m_i} | v \in f(w)\})$. By (P4), $M_i, \varepsilon \vDash p_i$ for all $\varepsilon \in c_i$. By ($\mathcal{R}$-monotonicity), there is a choice $c' \in \mathcal{C}(\{v^{\omega} | v \in f(w)\})$ such that if $M' = \langle W^P, W^I, f', V \rangle$, where $f'(w) = c'$, then $M', \varepsilon' \vDash p_i$ for all $\varepsilon' \in c'$. Since $\{p_1, \ldots, p_k\} \vdash^n_{\mathcal{R}} q$, there will be a choice $c'' \in \mathcal{C}(\{v^{\omega+n} | v \in f(w)\})$ such that if $M'' = \langle W^P, W^I, f'', V \rangle$, where $f''(w) = c''$, then $M'', \varepsilon'' \vDash q$ for all $\varepsilon'' \in c''$. By (P4), $M'', w \vDash Bq$. By ($n$-variation), $f'' \in \mathcal{F}^{\omega+n}(M, w)$, and, in turn, by ($n$-accessibility), $(M, w) \overset{\omega+n}{\sim} (M'', w)$. So for some model $(M'', w'') : (M, w) \overset{\omega+n}{\sim} (M'', w'')$, $M'', w'' \vDash Bq$. Thus, by (P5), $M, w \vDash \langle \omega + n \rangle Bq$. □

Theorem 1 says that if a conclusion $q$ follows within $n$ steps of reasoning from a set of premises $\{p_1, ..., p_k\}$, and the agent believes the $i$th premise after some $m_i$ steps of reasoning, for $1 \leq i \leq k$, then the agent believes $q$ after some $n + m_1 + m_2 + \cdots + m_k$ steps of reasoning. For instance, if an agent believes $p$ after 1 step and believes $p \to q$ after 2 steps, then she can apply modus ponens once to infer $q$ and hence believe $q$ after $1 + 2 + 1 = 4$ steps of reasoning.

The following result is a special case of theorem 1:

**Corollary 1.** ***($n$-distribution)*** *If* $\{p_1, \ldots, p_k\} \vdash^n_{\mathcal{R}} q$ *and* $Bp_i$*, for* $1 \leq i \leq k$*, then* $\langle n \rangle Bq$*.*

According to ($n$-distribution), if a conclusion $q$ follows within $n$ steps of reasoning from a set $\{p_1, \ldots, p_k\}$ of premises, and the agent believes all of the premises, then the agent believes $q$ after some chain of $n$-step reasoning. We can understand ($n$-distribution) as a dynamic counterpart of the distribution axiom **K** from standard doxastic logic:

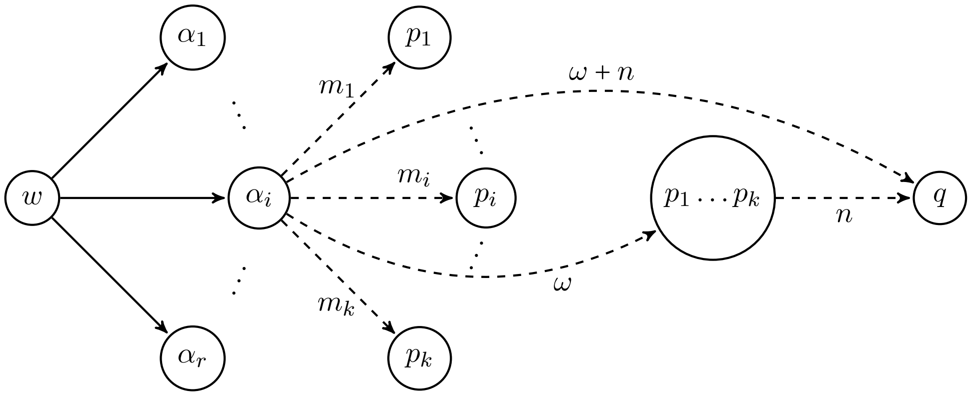


Figure 3: Diagrammatic representation of the proof of theorem 1. As in figure 2, the $p$s and $q$s are not names of worlds, but indicate that $p$ is true at the relevant worlds.

**K** $(Bp \wedge B(p \to q)) \to Bq$.

While **K** entails that beliefs are closed under believed entailment, ($n$-distribution) carries no such commitment: it merely says that agents have the ability to immediately form a belief in any proposition that follows within $n$ steps of reasoning from what they already believe.

As a special case of ($n$-distribution), we get:

**Corollary 2.** ***($n$-necessitation)*** *If* $\vdash^n_{\mathcal{R}} p$*, then* $\vDash \langle n \rangle Bp$.

According to ($n$-necessitation), if $p$ follows from the empty set within $n$ steps of reasoning, then the agent believes $p$ after some chain of $n$-step reasoning. We can see ($n$-necessitation) as a dynamic counterpart of the necessitation rule from standard doxastic logic:

**Nec** If $\vdash p$, then $\vdash Bp$.

While **Nec** entails that agents believe all logical truths, ($n$-necessitation) carries no such commitment: it merely says that agents have the ability to immediately form a belief in any proposition that is $n$-step inferable in $\mathcal{R}$ from the empty set.

We are now in a position to show how our framework successfully models agents who are both logically non-omniscient and logically competent. To see how logical omniscience is avoided, suppose that $Bp$ is true at $w$, for some $w \in W^P$, and consider any $q$ that is logically entailed by $p$. By (P4), $p$ is true at all doxastically accessible worlds for the agent. Since we quantify over both possible and impossible worlds in (P4), however, $q$ need not be true at all these doxastically accessible worlds. So $Bq$ need not be true at $w$, and logical omniscience is avoided. Also, our semantics for $\langle n \rangle Bp$ does not commit us to claiming that agents have an ability to immediately tease out all logical consequences of what they believe. In general, if $Bp$ is true at $w$, and $q$ follows from $p$ in more than $n$ steps of logical reasoning, it is easily verified that $\langle n \rangle Bq$ is false at $w$. Hence there are plenty of non-trivial logical consequences that agents cannot immediately infer.

To see how logical competence is secured, suppose $Bp$ is true at $w$, for some $w \in W^P$, and consider any $q$ that follows from $p$ within $n$ steps of logical reasoning. By ($n$-distribution), $\langle n \rangle Bq$ is true at $w$, and so it follows that the agent has an ability to immediately form a belief in any trivial logical consequence of what she already believes. So our model explains why the agent need never miss out on anything trivial: she can always make any trivial consequence $q$ of her beliefs count in reasoning and action. In turn, ($n$-distribution) helps us explain how logically competent agents pass the behavioral test. For suppose we ask the agent whether $q$ is the case. While $Bq$ need not be true at $w$—a desirable result in light of the collapse result—the fact that $\langle n \rangle Bq$ is true at $w$ tells us that the agent can immediately form a belief in $q$ and answer "yes" to the question on the basis of that belief.

By varying the value of $n$, our framework can model a whole *spectrum* of agents with different computational resources available for trivial logical reasoning. In the limit where $n = 0$, agents have no available computational resources, and $\langle n \rangle Bq$ will be false for any $q$ that follows from the agent's beliefs—assuming, as above, that nothing is 0-step inferable in $\mathcal{R}$. In the opposite limit, where $n$ goes towards infinity, agents have unlimited computational resources, and, by (Corollary 1), $\langle n \rangle Bq$ will be true for any $q$ that follows logically from the agent's beliefs. In between these extremes, we find agents with an intermediate amount of resources available for trivial logical reasoning. For such agents, as we have seen, $\langle n \rangle Bq$ will be true for some, but not all logical consequences $q$ of what they believe.[11]

It is worth noting that if we identify $\mathcal{R}$ with a complete proof system for propositional logic, and if we let $n$ approach infinity, we get the following pleasant symmetry between ($n$-distribution) and the axiom **K**: for any $q$ that is a classical logical consequence of an agent's beliefs, $\langle n \rangle Bq$ will be true in our dynamic doxastic logic just in case $Bq$ is true in standard doxastic logic. For $q$ will then be $n$-step inferable in $\mathcal{R}$ whenever $q$ is inferable (simpliciter) in propositional logic. In this sense, our framework can model agents who—much like logically omniscient agents—can trivially or immediately tease out all logical consequences of what they believe, including all logical truths.[12]

## 4 Jago's response to the collapse result

Mark Jago (2013; 2014) has recently proposed a framework that, much like ours, promises to steer clear of both logical omniscience and logical incompetence. In this section, we argue that Jago's framework does not hold its promise and hence that our framework is superior.

In light of the collapse result, Jago rightly concludes that agents must either be logically omniscient or fail to believe at least some trivial consequences of what they believe—he refers to this dilemma as "the problem of rational knowledge" (Jago 2013, p. 1152). On pain of logical omniscience, Jago accepts that it is "possible to fail to know or believe trivial truths (and, more generally, trivial consequences of one's beliefs)" (Jago 2014, p. 243). As he notes,

> [t]here is, to be sure, something counter-intuitive in this result. If an agent fails to know or believe that $A \vee \neg A$ for some '$A$', then her epistemic or doxastic state misses out on something trivial. Similarly, if an agent knows or believes that such-and-such, from which '$A$' trivially follows and yet she does not know or believe that $A$, then again

[11] While we have focused on logically competent agents who can infer at least the *trivial* logical consequences of what they believe, note that our framework is general enough to model what agents can come to believe after *any* finite chain of step-based reasoning, whether trivial or not.

[12] Thanks to an anonymous reviewer for drawing our attention to this point.

> her epistemic or doxastic state misses out on something trivial. I'll call such cases, in which an agent fails to know (or believe) some trivial consequence of what she knows (or believes), *epistemic oversights*. [...] Epistemic oversights are bizarre, but we know they must exist. For every logically non-omniscient agent, there is some knowledge she has which trivially entails something she does not know. Otherwise, her knowledge would be closed under all trivial inference rules and hence deductively closed. (Jago 2014, p. 206)

Epistemic oversights are "counter-intuitive" or "bizarre", according to Jago, because an agent who suffers from an epistemic oversight seems to "miss out on something trivial" and so seems irrational or logically incompetent.[13] To avoid treating agents as logically incompetent, Jago suggests that epistemic oversights must always be *indeterminate*: we can never rationally assert that an agent has a particular epistemic oversight. For if we do, we thereby treat her as logically incompetent (Jago 2013, p. 1152).

To ensure that epistemic oversights are always indeterminate, Jago develops epistemic models that satisfy the following principle:

> **(Indeterminacy)** [I]f '$A$' is a trivial consequence of what an agent $i$ knows, then it's never determinate that $i$ fails to know that $A$. (Jago 2013, pp. 1166–1167)

According to Jago, since it is never rational to assert what is indeterminate, (Indeterminacy) prevents us from attributing particular epistemic oversights to agents and thereby prevents us from treating them as logically incompetent.

While we agree with Jago that logically non-omniscient agents must suffer from epistemic oversights—the collapse result shows that much—we believe that there are several problems involved in using (Indeterminacy) to avoid logical incompetence. Below we raise five such problems.

*1. (Indeterminacy) is dispensable.* We do not agree with Jago's claim that agents who suffer from determinate epistemic oversights must be logi-

[13] Since Jago's use of "rationality" and our use of "logical competence" are supposed to do more or less the same work, we will use these terms interchangeably in what follows.

cally incompetent. For even such agents can have the ability to infer the trivial consequences of what they believe. And such agents, as we have shown, need not "miss out on anything trivial": they can always make any trivial consequence of what they believe count in reasoning and action. Hence (Indeterminacy) is dispensable for modeling agents who are both logically non-omniscient and logically competent.

*2. (Indeterminacy) lacks independent motivation.* Jago might grant that (Indeterminacy) is dispensable but hold that it nevertheless does the required job: it allows us to treat agents as logically competent. In point 4 below we argue that (Indeterminacy) in fact cannot do this job, but even if it could, we can ask for some *independent* reasons to accept (Indeterminacy)—reasons, that is, that do not derive from the need to avoid logical incompetence. Jago suggests that (Indeterminacy) is motivated by a structural similarity between the problem of rational knowledge and the sorites paradox:

> The problem of rational knowledge can be formulated in terms of a step-by-step deduction $D$ from premises $C$ which the agent in question clearly knows, to a conclusion '$A$' that the agent clearly does not know. By assumption, not every step of reasoning in $D$ preserves the agent's knowledge (since we eventually arrive at a conclusion the agent does not know). Yet any attempt to say precisely at which point in the deduction the agent's knowledge gives out is doomed to failure. [...] Formulating the problem in this way brings out its structural similarity with the sorites paradox. In this case, the principle that rational agents know the trivial consequences of what they know plays the role that tolerance conditionals play in the sorites. The tolerance conditionals for 'red', for example, say that (in a sorites series of colour samples), if sample $n$ is red then so is sample $n + 1$. Clearly, not all such conditionals are true; but we cannot say or discover which is false. (Jago 2013, p. 1155)

However, the alleged structural similarity between the problem of rational knowledge and the sorites paradox *presupposes* rather than independently motivates (Indeterminacy). The similarity is brought out by the claim that "any attempt to say precisely at which point in the deduction the agent's knowledge gives out is doomed to failure" and the claim that "any attempt to say precisely which tolerance conditional fails to hold is doomed to failure." But the former claim is plausible only if we already accept (Indeterminacy). For if we do not, nothing prevents us from pointing out precisely at which point in a deduction the agent's knowledge gives out. So there is only a structural similarity between the problem of rational knowledge and the sorites paradox if (Indeterminacy) is already assumed. As such, Jago has offered no independent reasons to accept (Indeterminacy).

*3. (Indeterminacy) faces potential counterexamples.* Consider a simple, logically non-omniscient artificial agent who believes just the propositions $p_1, p_1 \to p_2, p_2 \to p_3, \ldots, p_{k-1} \to p_k$, and suppose the agent can apply modus ponens only once. Given this, there is exactly *one* trivial consequence of the agent's beliefs, namely $p_2$. Since the agent is logically non-omniscient, the collapse result shows that the agent must suffer from at least *one* epistemic oversight. And since $p_2$ is the *only* trivial consequence of the agent's beliefs, it follows that the agent cannot believe $p_2$. So we determinately know that the agent fails to believe $p_2$. But this runs counter to (Indeterminacy): logically competent agents may well suffer from determinate epistemic oversights.

*4. (Indeterminacy) lacks explanatory power.* Even if we set aside the problems above, we doubt that (Indeterminacy) can adequately capture the sense in which agents are logically competent. Consider again the simple agent from point 3. Due to her logical competence, the agent can be assumed to pass the behavioral test: she immediately answers "yes" when asked whether $p_2$ is true. On the face of it, this behavior seems hard to reconcile with the fact that the agent does not believe $p_2$. For how can the agent give the correct answer, when she does not believe the answer? As we have seen, our model provides a straightforward explanation: since the agent believes $p_1$ and $p_1 \to p_2$, ($n$-distribution) tells us that the agent believes $p_2$ after one step of reasoning in $\mathcal{R}$, where $\mathcal{R}$ in this case contains only modus ponens. And since, by description, the agent can immediately apply modus ponens once, she can immediately infer $p_2$ and base her positive answer on that belief. By contrast, (Indeterminacy) merely tells us that the agent's lack of belief in $p_2$ is *indeterminate*. We have already argued against this claim, but even if we grant it, we are still not offered an explanation of why the agent assents to $p_2$ despite not believing it.

Of course, the agent above is quite simple: her reasoning mechanism is highly incomplete, her computational resources are very limited, and she only has a small number of beliefs. One might wonder how (Indeterminacy) fares in more complex cases where there are more than one trivial consequence of an agent's beliefs. Consider a logically non-omniscient, yet logically competent agent such that there are $m$ different trivial consequences $q_1, \ldots, q_m$ of the agent's beliefs. Suppose now that we ask this agent a series of questions concerning these trivial consequences. First, we ask her whether $q_1$ is true, then whether $q_2$ is true, and so on. Due to her logical competence, we can assume that she immediately answers "yes" to each of these questions. Due to her logical non-omniscience, however, we know that she must suffer from at least *one* epistemic oversight. So again we face an explanatory challenge: how can the agent answer "yes" to each question when she fails to believe at least *one* of the answers? As above, our model gives a straightforward answer: when asked about $q_i$, ($n$-distribution) says that the agent can immediately infer $q_i$ and base her answer on that derived belief. (Indeterminacy), by contrast, does not help us explain the agent's logically competent behavior. It prevents us from rationally ascribing any particular epistemic oversight to the agent, but it does not explain how the agent can answer correctly each question despite failing to believe at least one of the answers.

*5. Elusive justification for belief ascriptions.* The case above also points out a different tension in Jago's proposal. When the agent answers "yes" to the question concerning $q_1$, we seem to acquire good (albeit fallible) justification for saying that the agent believes $q_1$. By repeating this procedure for $q_2, \ldots, q_m$, we seem to acquire good justification for saying that the agent believes each $q_i$. Yet, due to her logical non-omniscience, we determinately *know* that the agent fails to believe at least *one* $q_i$. So we seem to end up with good justification for the claim that the agent believes each of $q_1, \ldots, q_m$ and yet fails to believe at least one of $q_1, \ldots, q_m$. To avoid this absurdity, it seems that Jago must hold that we somehow lose justification for at least one of the belief ascriptions during the process of questioning. It is unclear to us why our justification for belief ascriptions should be "elusive" in this way. But if it is, we should expect an independently plausible story of why our justification is lost during the process of questioning—a story that Jago does not give.

In sum, since our dynamic framework does not appeal to (Indeterminacy), it avoids the problems and tensions that Jago faces. Thus, we claim, our framework provides a superior model of agents who are both logically non-omniscient and logically competent.

# 5 Conclusion

We have claimed that a solution to the problem of logical omniscience should allow us to model agents who are both logically non-omniscient and logically competent. While standard versions of the impossible worlds model of belief aim to describe such agents by imposing closure constraints on impossible worlds, the collapse result shows that even very weak such constraints lead to logical omniscience. Instead we have developed a dynamic impossible worlds framework and argued that it successfully avoids the Scylla of logical omniscience and the Charybdis of logical incompetence.

# References

[Author] (2013). In:

— (2014). In:

— (2015). In:

— (2016). In:

Berto, F. (2013). "Impossible Worlds". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2013.

Chalmers, D. (2011). "The Nature of Epistemic Space". In: *Epistemic Modality*. Ed. by A. Egan and B. Weatherson. New York: Oxford University Press, pp. 60–107.

Cherniak, C. (1986). *Minimal Rationality*. MIT Press.

Cresswell, M. (1970). "Classical Intensional Logic". In: *Theoria* 36, pp. 347–372.

— (1972). "Intensional Logics and Logical Truth". In: *Journal of Philosophical Logic* 1, pp. 2–15.

— (1973). *Logics and Languages*. Methuen and Co.

Ditmarsch, W., W. Hoek, and B. Kooi (2008). *Dynamic Epistemic Logic*. Springer.

Drapkin, J. and D. Perlis (1986). "A Preliminary Excursion into Step-Logics". In: *Proceedings of the SIGART International Symposium on Methodologies for Intelligent Systems*. Ed. by C. Ghidini, P. Giodini, and W. van der Hoek, pp. 262–269.

— (1990). "Reasoning Situated in Time I: Basic Concepts". In: *Journal of Experimental and Theoretical Artificial Intelligence* 2, pp. 75–98.

Duc, H.N. (1997). "Reasoning about rational, but not logically omniscient, agents". In: *Journal of Logic and Computation* 7.5, pp. 633–648.

Elgot-Drapkin, J. et al. (1999). *Active Logics: A Unified Formal Approach to Episodic Reasoning*. Tech. rep. University of Maryland.

Fagin, R., J. Halpern, and M. Vardi (1995). "A Nonstandard Approach to the Logical Omniscience Problem". In: *Artificial Intelligence* 79, pp. 203–240.

Fagin, R. et al. (1995). *Reasoning About Knowledge*. MIT Press.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Two Notions*. Cornell University Press.

— (1975). “Impossible Possible Worlds Vindicated”. In: *Journal of Philosophical Logic* 4, pp. 475–484.

Jago, M. (2013). “The Problem of Rational Knowledge”. In: *Erkenntnis*, pp. 1–18.

— (2014). *The Impossible: An Essay on Hyperintensionality*. Oxford University Press.

Konolige, K. (1986). *A Deduction Model of Belief*. Morgan Kaufman.

Lakemeyer, G. (1987). “Tractable Meta-Reasoning in Propositional Logics of Belief”. In: *Tenth International Joint Conference on Artificial Intelligence*, pp. 198–202.

Levesque, H.J. (1984). “A Logic of Implicit and Explicit Belief”. In: *National Conference on Artificial Intelligence*, pp. 198–202.

Rantala, V. (1982). “Impossible Worlds Semantics and Logical Omniscience”. In: *Acta Philosophica Fennica* 35, pp. 106–115.

Wansing, H. (1990). “A General Possible Worlds Framework for Reasoning about Knowledge and Belief”. In: *Studia Logica: An International Journal for Symbolic Logic* 49, pp. 523–39.

Weirich, P. (2004). *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. OUP.