AN ALGORITHMIC IMPOSSIBLE-WORLDS MODEL OF BELIEF AND KNOWLEDGE

ZEYNEP SOYSAL
Department of Philosophy, University of Rochester

Abstract. In this paper, I develop an algorithmic impossible-worlds model of belief and knowledge that provides a middle ground between models that entail that everyone is logically omniscient and those that are compatible with even the most egregious kinds of logical incompetence. In outline, the model entails that an agent believes (knows) ϕ just in case she can easily (and correctly) compute that ϕ is true and thus has the capacity to make her actions depend on whether ϕ. The model thereby captures the standard view that belief and knowledge ground are constitutively connected to dispositions to act. As I explain, the model improves upon standard algorithmic models developed by Parikh, Halpern, Moses, Vardi, and Duc, among other ways, by integrating them into an impossible-worlds framework. The model also avoids some important disadvantages of recent candidate middle-ground models based on dynamic epistemic logic or step logic, and it can subsume their most important advantages.

§1. Introduction. According to the standard possible-worlds models of belief and knowledge, a person, S, believes (knows) a proposition, ϕ, if and only if ϕ is true in all the possible worlds that are doxastically (epistemically) accessible to S. These models have the following consequence:

> *Full Logical Omniscience:* If S believes (knows) all the propositions in set Φ, and Φ logically entails ψ, then S believes (knows) ψ.

The problem of logical omniscience for the standard model is that Full Logical Omniscience is clearly false: everyone fails to believe some logical consequences of the propositions that they believe, and everyone fails to believe some logical truths. The standard model thus at best provides an idealization of the notions of belief and knowledge.

At the opposite end of the spectrum are models without any logical constraints on belief and knowledge, such as impossible-worlds models with a maximally permissive construal of the impossible worlds. On impossible-worlds models, S believes (knows) ϕ if and only if ϕ is true in all the possible *or impossible* worlds that are doxastically (epistemically) accessible to S. On a maximally permissive construal of the impossible worlds, there are impossible worlds in which, for instance, $\phi \wedge \psi$ is true but neither ϕ nor ψ is true. If such a world is doxastically accessible for some person, S, then S can believe $\phi \wedge \psi$ without believing either ϕ or ψ. The advantage of models of belief and knowledge that have no logical constraints is that for any given constraint, it

Received: April 5, 2022.
2020 *Mathematics Subject Classification*: Primary 03B42.
Keywords and phrases: algorithmic knowledge, problem of logical omniscience, impossible worlds, doxastic logic, epistemic logic, bounded rationality.

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Association for Symbolic Logic. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

doi:10.1017/S1755020323000059

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

seems possible for there to be someone who violates it: borrowing an example from Nolan [33, p. 47], someone convinced that a god is beyond logic might believe that their god exists and doesn't exist, while at the same time not believing that their god doesn't exist. Arguably, a model of belief shouldn't by fiat rule out the possibility of such (albeit unusual) individuals. But these maximally permissive models also have the disadvantage that they don't satisfy three common desiderata for models of belief and knowledge.

The first of these desiderata is to capture ordinary agents who are logically non-omniscient but still logically competent. An agent is logically competent when, for instance, she "know[s] at least a (sufficiently) large class of logical truths, and can draw sufficiently many conclusions from their knowledge" [13, p. 241] or "she at least does not miss out on any *trivial* logical consequences of what she believes" [8, pp. 502f.], where what counts as "sufficiently many" or "trivial" logical consequences could depend on the agent's computational capacities and thus be agent-relative (see [8, p. 503; 14, p. 639]). If one's goal is to model agents who are in such senses logically "competent" but non-omniscient, then one needs a middle ground between models that entail logical omniscience and those that leave open complete logical incompetence.[1]

The second desideratum is to capture constraints that the nature of logical concepts (arguably) imposes on logically related beliefs. For instance, some have suggested, contra Nolan's example, that possessing the concept of conjunction requires believing ψ when one believes $\phi \wedge \psi$ (e.g., Jago [23, pp. 163–169; 24, pp. 1151f.]), or at least being disposed, when certain normal conditions are in place, to believe ψ if one believes $\phi \wedge \psi$ (e.g., Boghossian [9, pp. 493–497], Warren [55, pp. 46f.]). Maximally permissive models leave open that one can believe or fail to believe any combination of logically related beliefs, and thus aren't useful if one's aim is to capture (apparent) constitutive constraints on logically related beliefs.[2]

The third, and in my view the most important, desideratum is to capture that whatever an agent can computationally "easily access" is—by the very nature of belief and knowledge—already part of what she believes or knows.[3] On the most common understanding, belief is, or at least grounds, a certain class of dispositions to act.[4] As the standard example goes, one believes that there is beer in the fridge just in case one is disposed to go to the fridge if one wants to drink beer, to answer "Yes" to the question of whether there is beer in the fridge if one wants to be truthful, and so on. For dispositionalists or functionalists about belief such as Lewis [29, 30] or Stalnaker [47], having certain dispositions to act is even partly constitutive of what it is to have a belief. On Stalnaker's view:

> To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true. [47, p. 15]

[1] Jago [24, p. 1152], Skipper [38, pp. 3f.], Solaki [41, p. 2], and Solaki, Berto, and Smets [43, p. 740] motivate the need for a middle-ground model in this way.

[2] Jago [23, pp. 163–169] motivates the need for a middle-ground model in this way.

[3] Note that this is distinct from the desideratum that whatever an agent can computationally easily access *from the propositions that she believes or knows* is already part of what she believes or knows. I discuss the latter and its contrast with my third desideratum in Section 2.2.

[4] For explanation of and literature about this standard understanding, see, e.g., [37].

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

Knowledge, in turn, is on this view a certain kind of capacity: as Stalnaker puts it, "[k]nowledge whether ϕ [...] is the capacity to make one's actions depend on whether ϕ" [51, pp. 2f.]. On any view on which belief and knowledge ground or are constitutively connected to behavioral dispositions, propositions that are computationally "easily accessible" to an agent should already be part of what she believes or knows. This is because whenever some information is only a trivial computation or inference away for an agent, the agent already has the capacity to act upon that information. For instance, assume that Ola knows that adding 2 to a number $d_0 \dots d_n 2$ yields $d_0 \dots d_n 4$: she is able to answer relevant questions correctly; she uses this information in ordinary life, such as in calculating tips, buying the right amount of certain things, and so on. Assume, further, that Ola has never explicitly thought about the number $19,822$, but that she is immediately able to give the correct answer to questions such as "What is $19,822 + 2$?" or "Is $19,822 + 2 = 19,824$?" It would then seem that Ola also knows that $19,822 + 2 = 19,824$. But the dispositions or capacities characterizing this knowledge involve a short computation: Ola can only manifest her disposition to answer questions about $19,822 + 2 = 19,824$ correctly after, for instance, replacing x in "$x2 + 2 = x4$" with "$19,82$." On the standard dispositional understanding of belief and knowledge, this doesn't mean that Ola didn't know that $19,822 + 2 = 19,824$ until she performed a calculation or inference. Rather, using terminology from Stalnaker [48, pp. 435f., 439], what Ola believes or knows is what is "available" to guide her behavior, even if it hasn't ever been used or "accessed" to do so. Since maximally permissive models allow that one doesn't believe what one can computationally "easily access," they are unable to capture the dispositional nature of belief and knowledge. Models that entail logical omniscience, too, seem unable to capture the dispositional nature of belief and knowledge. As Stalnaker notes:

> Because of our computational limitations, we may have the capacity constituted by the knowledge that P, or the disposition constituted by the belief that P, while at the same time lacking the capacity or disposition that we would have if we knew or believed some deductive consequence of P. [48, p. 436]

For instance, it seems clear that Ola could be disposed to make her actions depend on whether the number of students at her school is $6,299$, but lack the disposition to make her actions depend on whether the number of students at her school is prime: she could order the right amount of school gear, answer questions about the number of students correctly, and so on, while being at a loss when asked whether the number of students at her school is prime.[5] The standard dispositional understanding of belief and knowledge thus seems to require a middle-ground model: not all the logical consequences of one's beliefs or knowledge are accessible to guide action and thus are believed or known, but those pieces of information that are easily accessible are already believed or known.[6]

My aim in this paper is to develop a middle-ground model of belief and knowledge that satisfies these three desiderata. My proposal builds upon algorithmic models

[5] As I explain in Section 2.2, Stalnaker's own formulation of the functionalist definition of belief in [47, p. 15] entails Full Logical Omniscience in a possible-worlds framework, but not in an impossible-worlds framework.

[6] Bjerring and Skipper [8, p. 503] also mention this desideratum, but, as I explain in Section 3, their view doesn't satisfy it.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

developed by Parikh [34]; Halpern, Moses, and Vardi [17]; and Duc [15]. The unifying idea of algorithmic models is that whether an agent believes or knows something depends on the agent's internal algorithms, and thus on her computational capacities. As I will explain, algorithmic models are thereby particularly well-suited to capture the idea that an agent already believes or knows what is easily computationally accessible to her, to model logically competent but non-omniscient agents, and to account for possible constitutive relations between logically related beliefs. Algorithmic models have been developed and studied in logic and computer science, and have been applied in the study of security protocols and cryptography.[7] But there has been very little discussion or development of algorithmic models in the philosophical literature on the problem of logical omniscience.[8] My aim here is to fill this gap by developing an algorithmic model that satisfies the philosophical desiderata for middle-ground models better than any existing algorithmic model, and by motivating it philosophically. I end in Section 3 by comparing the algorithmic strategy for developing a middle-ground model to approaches that use dynamic epistemic logic or step logic, such as the one developed recently by Bjerring and Skipper [8].

§2. Algorithmic models. Algorithmic models of belief and knowledge are developed as alternatives to the standard possible- and impossible-worlds models to solve the problem of logical omniscience. The guiding idea behind algorithmic models is that belief and knowledge have a computational aspect that isn't captured by the standard possible-worlds models or by the standard approaches to solving the problem of logical omniscience.[9] Although they don't explicitly say so, Parikh [34] and Halpern et al. [17] motivate this guiding idea from the type of functionalist perspective, outlined in Section 1, on which knowledge is a certain kind of capacity to act. For instance:

> We have tried in this paper to make a case that real knowledge is not a *set* but a *behaviour*. [34, p. 7]
>
> [A]n agent that has to act on his knowledge has to be able to compute this knowledge; we do need to take into account the algorithms available to the agent, as well as the "effort" required to compute knowledge. [17, p. 256]

Different algorithmic models give different interpretations to this guiding idea and to what it means to "compute knowledge." But they all share the assumptions that agents have algorithms, and that whether an agent knows or believes ϕ depends on her algorithm and its output when given ϕ. On one way of putting it in intuitive terms, the view is that one believes ϕ just in case one's algorithm efficiently computes that ϕ is

[7] Recent applications of algorithmic models include [20]; Fagin, Halpern, Moses, and Vardi [16, pp. 412f.] list older applications.

[8] Exceptions include [27, 34, 44], and mentions in literature surveys in [22, pp. 95–97; 40, p. 25].

[9] For a survey of standard approaches to solving the problem of logical omniscience, see [16, chap. 9]. As Halpern et al. [17, p. 261], Fagin et al. [16, pp. 398f.], and Halpern and Pucella [19, p. 231] explain, although algorithmic models differ from these other approaches by explicitly modeling agents as having algorithms, they can subsume some of them, such as awareness or syntactic approaches, with the addition of certain assumptions about the relevant algorithms or about the notion of awareness.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

true, and one knows ϕ just in case one's algorithm efficiently and correctly computes that ϕ is true.[10] Failures of logical omniscience are then diagnosed as computational failures: if an agent knows or believes ϕ but fails to know or believe some ψ entailed by ϕ, this is because the agent doesn't have either the right kind of algorithm or sufficient computational resources. This fits with an intuitive characterization of cases of failures of logical omniscience: for instance, if Ola knows that the number of students is 6,299 but not that the number of students is prime, or if she knows the axioms of number theory but not some theorem, this is because she can't check for primality or theoremhood, she can't retrieve this information from a reliably stored memory base, and so on, or because it would take her too long to run such algorithms.[11]

Already at this level of generality, algorithmic models should strike us as promising starting points for developing a middle-ground model of belief and knowledge. On algorithmic models, an agent knows whatever her algorithm can efficiently compute. One can thus hope to delineate logically competent agents by putting certain constraints on the algorithms and resources that these agents use. For instance, logically competent agents might have algorithms that compute a certain class of logical truths with minimal effort. Similarly, one can hope to model constitutive connections between logically related beliefs by putting constraints on the kinds of algorithms that conceptually competent agents use. For instance, being conceptually competent with conjunction might require having an algorithm that computes that ψ is true if it computes that $\phi \wedge \psi$ is true. Finally, algorithmic models capture the view that belief and knowledge are connected to action. In general, it is highly plausible that whenever one is disposed to exhibit some behavior, this is because one has an (internal) algorithm that produces this behavior. It thus makes sense to model agents as having algorithms, and to model belief and knowledge as dependent on the characteristics of these algorithms. Algorithmic models thereby also straightforwardly capture the view that whatever one is able to efficiently compute and thus act upon is already part of what one believes (knows), because they entail that efficiently (and correctly) computing that ϕ is true is sufficient for believing (knowing) ϕ.

One disadvantage of the existing algorithmic models that we can already see is that they give up on the worlds-based framework for modeling belief and knowledge: one's belief and information states are no longer modeled as sets of worlds, belief and knowledge are no longer modeled as truth in all doxastically or epistemically accessible worlds, and acquiring beliefs and learning are no longer modeled as ruling out doxastic or epistemic possibilities. This is unfortunate, for the worlds-based framework captures some important aspects of belief and knowledge—such as their independence from linguistic action and world-connectedness—in a formally elegant manner, and it yields a unified account of mental, linguistic, and informational content.[12] In Section 2.1, I outline the formal details of the existing algorithmic models that are most relevant given our purposes and explain their other advantages and disadvantages. In Section 2.2,

[10] This is roughly how Parikh [34, p. 5] and Halpern and Pucella [19, p. 222] put it. As we will see in Section 2.1, Halpern et al. [17] instead say that the algorithm computes that ϕ is true in all accessible worlds.

[11] See [19, 34, 48–50, 52] for the claim that failures of logical omniscience are intuitively characterized as computational failures.

[12] Stalnaker [46, 47], Lewis [30, sec. 1.4], Nolan [32], Jago [23, pp. 24–27], and Berto and Jago [5, pp. 213–216] explain the benefits of the worlds-based framework for modeling belief and knowledge.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

I then develop an algorithmic model that satisfies the desiderata for a middle-ground model and doesn't have these disadvantages.

2.1. Standard algorithmic models. The standard and most sophisticated existing algorithmic model is due to Halpern et al. [17]. In the single-agent and static setting (i.e., where we don't consider the evolution of belief and knowledge over time), Halpern and Pucella [19] work with standard Kripke structures of the form $\langle \mathcal{W}, \mathcal{W}', \pi \rangle$, where $\mathcal{W}$ is a set of possible worlds, $\mathcal{W}'$ is the set of possible worlds that are accessible to the agent, and π is an interpretation function that associates each possible world $w \in \mathcal{W}$ with a truth assignment $\pi(w)$ to the primitive propositions of the language.[13] An *algorithmic knowledge structure* is defined as a tuple $\mathcal{M} = \langle \mathcal{W}, \mathcal{W}', \pi, \mathcal{A} \rangle$ where $\langle \mathcal{W}, \mathcal{W}', \pi \rangle$ is a Kripke structure and $\mathcal{A}$ is a *knowledge algorithm* that takes as input a formula and returns either "Yes," "No," or "?" (knowledge algorithms are thus assumed to terminate). "The agent knows ϕ," symbolized standardly as "$K\phi$," then gets the following satisfaction (or truth) conditions:

$$\langle \mathcal{M}, w \rangle \vDash K\phi \quad \Leftrightarrow \quad \mathcal{A}(\phi) = \text{“Yes.”}$$

To capture the evolution of knowledge over time, Halpern et al. [17] give a run-based semantics on which agents can use different algorithms in different states. The local state of an agent is modeled as consisting of some local data and a local algorithm. The agent then knows ϕ at a state if her local algorithm outputs "Yes" when given both ϕ and the local data as inputs.[14]

On this semantics, whether one knows ϕ has nothing to do with the truth-value of ϕ in any accessible world, and there are no constraints on the algorithm $\mathcal{A}$. Halpern and Pucella [19, p. 223] propose the following constraint: a knowledge algorithm $\mathcal{A}$ is *sound* for $\mathcal{M}$ if and only if for all ϕ, $\mathcal{A}(\phi) =$ "Yes" implies $\langle \mathcal{M}, w \rangle \vDash \phi$ for all $w \in \mathcal{W}'$ and $\mathcal{A}(\phi) =$ "No" implies $\langle \mathcal{M}, w \rangle \vDash \neg\phi$ for some $w \in \mathcal{W}'$.[15] That is, an algorithm is sound just in case an output of "Yes" implies that ϕ true in all epistemically accessible worlds (and thus "known" in the sense of the standard possible-worlds model), while an output of "No" implies that ϕ is false in some epistemically accessible world (and thus "unknown" in the standard model's sense). This is the sense in which knowledge algorithms "compute knowledge" on Halpern et al.'s [17] construal: they compute whether the input formula is true in all accessible worlds. According to Halpern et al. [17, p. 259], what is defined without the soundness constraint is a notion of belief, i.e., using "$B\phi$" to formalize "the agent believes ϕ":

$$\langle \mathcal{M}, w \rangle \vDash B\phi \quad \Leftrightarrow \quad \mathcal{A}(\phi) = \text{“Yes.”}$$

Knowledge is thus defined as above but with the additional constraint that $\mathcal{M}$'s algorithm is sound.

The standard algorithmic model avoids the problem of logical omniscience: since there are no constraints on the knowledge algorithms, one's knowledge algorithm can output "Yes" given ϕ but either "No" or "?" given ψ even if ϕ entails ψ or is logically equivalent to it. But given that one's knowledge algorithms can be extremely weak, we don't yet have a model that captures agents who are logically (or conceptually)

[13] In [19] this is assumed to be a K45 Kripke structure.
[14] For details, see [17, pp. 257–260].
[15] In the run-based semantics, an algorithm is sound for an agent if and only if this holds at all points in which the algorithm is local [17, p. 259].

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

competent. For instance, an agent can have a sound knowledge algorithm that never outputs "Yes." As it stands, the standard algorithmic model thus isn't an adequate middle-ground model.

There are other disadvantages of the standard algorithmic model given our purposes. The first is that the model is overly linguistic. Belief and knowledge are assumed to manifest always and only in linguistic action: the agent is given a sentence and in return provides a verbal response. But this is too narrow, as people can have knowledge that they aren't able to verbally articulate. Stalnaker provides many such examples: the "shrewd but inarticulate" chess player who can access information for choosing a move but not for answering questions [48, p. 439], or the experienced outfielder who knows exactly when and where the ball will come down for the purpose of catching the ball but not for the purpose of answering the question "Exactly when and where is the ball going to come down?" [50, p. 263]. More generally, on the dispositional understanding, belief and knowledge are supposed to help explain people's overall behavior—whether linguistic or otherwise.

A related disadvantage of the algorithmic model is that the objects of belief and knowledge are taken to be sentences, because the algorithms operate on sentences. Parikh [35, pp. 472–474], whose algorithmic notion of knowledge is even called "linguistic knowledge" [34, p. 4], explains that this choice is made so that belief and knowledge on the resulting account aren't closed under necessary equivalence (if ϕ and ψ are possible-worlds propositions and necessarily equivalent, then $\phi = \psi$, and thus the algorithm's output is the same for ϕ and ψ.) As we will see in Section 2.2, one can maintain that the objects of belief and knowledge are propositions while avoiding closure under necessary equivalence by moving to an impossible-worlds framework.

Finally, the most important gap in the standard algorithmic model given our purposes is that it doesn't address limitations of computational resources. On this model, an agent knows ϕ if her knowledge algorithm will eventually output "Yes" given ϕ, but that could take an unlimited amount of computational resources. For our purposes, we need to model bounds on the resources that the algorithms can use: In particular, if it would take an agent too long to compute ϕ, then she is unable to act on the information that ϕ and thus neither knows nor believes ϕ. For instance, an outfielder who would need to sit down for 3 hours to calculate the trajectory of the ball clearly neither knows nor believes that the ball will come down at the relevant location and time before they perform the calculations.

Halpern et al. [17, pp. 260f.] briefly discuss this problem and mention that one could put constraints on local algorithms so that they have to complete their run within a given unit of time.[16] Duc [15, pp. 39–51] develops an algorithmic system for knowledge that involves this idea. He introduces the formula "$K_i^n\phi$" to stand for "if asked about ϕ, i is able to derive reliably the answer 'yes' within n units of time," and the formula "$K_i^{\exists}\phi$" to stand for "agent i can infer ϕ reliably in finite time" [15, p. 41]. The latter captures the spirit of Halpern et al.'s notion of knowledge: $K_i^{\exists}\phi$ holds just in case the agent has an algorithm that computes that ϕ is true within a finite but unbounded amount of time. In these definitions, the qualification "reliably" is supposed to imply

[16] They also note that if one wants to model algorithms with longer running times, the algorithms could be "split up among successive states" [17, p. 261]. In Section 3, I consider one way to develop this type of idea and explain why it doesn't yield an adequate middle-ground model.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

both that the agent's computation is correct (i.e., if either $K_i^n\phi$ or $K_i^{\exists}\phi$ hold, then ϕ is true) and that the agent doesn't choose a procedure that correctly computes ϕ by chance, but is able to "select deterministically a suitable procedure for the input" [15, p. 41]. The idea is that the agent has a general procedure or algorithm that, given ϕ, correctly computes that ϕ is true within n units of time, and this general algorithm might involve steps for choosing the right sub-algorithms to run on ϕ.

Duc [15] thus defines infinitely many "time-stamped" notions of knowledge, one for each time unit n. This is slightly counterintuitive, given that we presumably only have one non-time-stamped notion of knowledge. As I will explain in Section 2.2, however, it is highly plausible to define knowledge in terms of such time-stamped notions by using a threshold.

Duc [15, pp. 45–48] goes on to provide a derivation system for his language of algorithmic knowledge. Some of his assumptions are too strong for our purposes. He assumes that both $K_i^n\phi$ and $K_i^{\exists}\phi$ imply that the agent is able to prove ϕ. On this view, agents employ a decidable axiom system extending propositional logic, and thus "all proofs can be generated algorithmically" by a general-purpose theorem prover [15, p. 44]. Duc only considers agents who have such a general-purpose theorem prover. After analysing a query and trying out special algorithms, these agents revert to using the general-purpose theorem prover. Thus, if ϕ is a theorem of the agent's axiom system, the agent will eventually find its proof. In other words, the following rule of inference is valid in Duc's system:

$$K_i^{\exists}\phi \text{ may be inferred from } \phi. \qquad (\mathbf{NEC}^{\mathrm{A}})$$

Similarly, Duc assumes that if formulae $\phi_1, \ldots, \phi_n$ can all be derived in the agent's system and $\phi_1 \wedge \cdots \wedge \phi_n \rightarrow \psi$ is a theorem, then the agent will also eventually output a proof of ψ if queried about it [15, p. 44]. A special case of this principle is that agents can use modus ponens in their reasoning, which yields the following as an axiom in Duc's system:

$$K_i^{\exists}\phi \wedge K_i^{\exists}(\phi \rightarrow \psi) \rightarrow K_i^{\exists}\psi. \qquad (\mathbf{K}^{\mathrm{A}})$$

As I will argue in Section 2.2, principles such as ($\mathbf{NEC}^{\mathrm{A}}$) and ($\mathbf{K}^{\mathrm{A}}$) can form plausible constraints on logically or conceptually competent agents. But we should avoid the extremely strong assumption that knowing ϕ always requires being able to find a proof of ϕ in some derivation system. After all, an agent might know even mathematical or logical truths by retrieving them from a memory base that was reliably stored (for instance, via expert testimony), without having any ability to produce proofs. Generally speaking, it is overly restrictive to think of all belief and knowledge that ϕ in terms of features of a proof of ϕ in some derivation system. (I come back to this point in Section 3.)

2.2. An algorithmic impossible-worlds model. My proposal is to build an algorithmic model on the basis of an impossible-worlds model. An important advantage of impossible-worlds models is that they preserve some core aspects of the standard possible-worlds model, including the idea that one's belief and information states are sets of worlds, that belief and knowledge are truth in all doxastically or epistemically accessible worlds, and that acquiring beliefs and learning are ruling out of doxastic or epistemic possibilities. Moreover, impossible worlds allow for more fine-grained constructions of content: sentences that are true in all the same possible worlds

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

correspond to the same possible-worlds proposition, but most often (if not always) differ in truth-values at impossible worlds and thus correspond to different propositions construed as sets of possible *and impossible* worlds.[17] As such, the impossible-worlds framework will enable us to avoid three disadvantages of the standard algorithmic framework discussed above, viz., that it gives up the standard worlds-based framework for modeling belief and knowledge, that belief and knowledge are assumed to manifest always and only in linguistic action, and that the objects of belief and knowledge are sentences and not propositions. Impossible-worlds frameworks also have some disadvantages, for instance, because the account of content that they yield is extremely fine-grained.[18] But they are widely and increasingly used for solving problems of hyperintensionality in philosophy and, as we will see in Section 3, they are used in the most important competitor approaches to developing a middle-ground model.[19]

Let us turn to some of the details of impossible-worlds models. Let $\mathcal{L}$ be the language of our model, defined as follows:

$$At := p, q, r, \ldots$$
$$\phi ::= At \mid \neg\phi \mid (\phi \wedge \phi) \mid B\phi \mid K\phi,$$

where "$B\phi$" formalizes "the agent believes ϕ" and "$K\phi$" formalizes "the agent knows ϕ." An *impossible-worlds model* is a tuple $\mathcal{M} = \langle \mathcal{W}, \mathcal{P}, d, e, v \rangle$, where $\mathcal{W}$ is a non-empty set of *worlds*; $\mathcal{P} \subseteq \mathcal{W}$ is a non-empty set of *possible worlds* (thus $\mathcal{I} := \mathcal{W} \setminus \mathcal{P}$ is the set of *impossible worlds*); $d : \mathcal{W} \to 2^{\mathcal{W}}$ is a *doxastic accessibility* function that assigns each world $w \in \mathcal{W}$ to the set of worlds *doxastically accessible* from w; $e : \mathcal{W} \to 2^{\mathcal{W}}$ is an *epistemic accessibility* function that assigns each world $w \in \mathcal{W}$ to the set of worlds *epistemically accessible* from w; and v is a valuation function that maps each atomic sentence $p \in At$ and world $w \in \mathcal{P}$ to either 0 or 1, and maps each sentence $\phi \in \mathcal{L}$ and world $w \in \mathcal{I}$ to either 0 or 1.[20] Since knowledge entails belief, it is standard to assume that doxastically accessible worlds are also epistemically accessible, i.e., that $d(w) \subseteq e(w)$ for all $w \in \mathcal{W}$. I also assume throughout that worlds are *centered*, i.e., they are worlds with a marked individual at a time. As Lewis [30, pp. 27–30] explains, this should be assumed in all doxastic and epistemic models because agents can obviously have different beliefs and knowledge at different times.[21] Finally, I add to the *satisfaction* relation (written "$\vDash$" as usual) the *dissatisfaction* (or making false)

[17] Nolan [31, p. 563] discusses how to construct propositions in terms of possible and impossible worlds. For these and other advantages of impossible-worlds models, see [4, 5, 33].

[18] Stalnaker [49] and Bjerring and Schwarz [7] discuss this criticism. Bjerring and Schwarz [7, pp. 23–30] also argue that permissive impossible-worlds models fail to preserve some important features of the standard possible-worlds model, viz., the recursive semantic rules of possible-worlds semantics and the idea that worlds are maximally specific ways things might be.

[19] That being said, Soysal [44] and Kipper, Kocurek, and Soysal [27] also develop similar algorithmic models based on a possible-worlds framework.

[20] This presentation follows [8, pp. 509f.] with minor changes. See also [16, pp. 357–362], and originally [36].

[21] For further motivation, for using centered worlds, see [29]. It isn't completely obvious how to develop a theory of centered impossible worlds, but see [11] for the related development of centered epistemically possible worlds.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

relation, written "$\dashv$." If $w \in \mathcal{P}$, $\vDash$ and $\dashv$ are defined recursively:

$$\begin{aligned}
\langle \mathcal{M}, w\rangle \vDash p \quad &\Leftrightarrow \quad v(p, w) = 1,\\
\langle \mathcal{M}, w\rangle \vDash \neg\phi \quad &\Leftrightarrow \quad \langle \mathcal{M}, w\rangle \nvDash \phi,\\
\langle \mathcal{M}, w\rangle \vDash (\phi \wedge \psi) \quad &\Leftrightarrow \quad \langle \mathcal{M}, w\rangle \vDash \phi \text{ and } \langle \mathcal{M}, w\rangle \vDash \psi,\\
\langle \mathcal{M}, w\rangle \vDash B\phi \quad &\Leftrightarrow \quad \langle \mathcal{M}, w'\rangle \vDash \phi \text{ for all } w' \in d(w),\\
\langle \mathcal{M}, w\rangle \vDash K\phi \quad &\Leftrightarrow \quad \langle \mathcal{M}, w'\rangle \vDash \phi \text{ for all } w' \in e(w),\\
\langle \mathcal{M}, w\rangle \dashv \phi \quad &\Leftrightarrow \quad \langle \mathcal{M}, w\rangle \nvDash \phi.
\end{aligned}$$

If $w \in \mathcal{I}$, then $\vDash$ and $\dashv$ are defined as follows:

$$\begin{aligned}
\langle \mathcal{M}, w\rangle \vDash \phi \quad &\Leftrightarrow \quad v(\phi, w) = 1,\\
\langle \mathcal{M}, w\rangle \dashv \phi \quad &\Leftrightarrow \quad v(\neg\phi, w) = 1.
\end{aligned}$$

The idea here is that a sentence ϕ is dissatisfied (or false) at a world if and only if its negation $\neg\phi$ is satisfied (or true) at that world, for both possible and impossible worlds. Given a possible world $w \in \mathcal{P}$, only one of ϕ and $\neg\phi$ is satisfied at w and the other is dissatisfied at w. But this needn't be the case at impossible worlds: given an impossible world $w \in \mathcal{I}$, both ϕ and $\neg\phi$ could be satisfied at w (if $v(\phi, w) = 1$ and $v(\neg\phi, w) = 1$), and neither ϕ nor $\neg\phi$ could be satisfied (if $v(\phi, w) = 0$ and $v(\neg\phi, w) = 0$). Impossible worlds can thus be both inconsistent and incomplete entities. I further adopt the *maximally permissive* construal of impossible worlds on which for any incomplete and/or inconsistent set of sentences $\Gamma \subseteq \mathcal{L}$, there is an impossible world $w \in \mathcal{I}$ such that for all $\phi \in \mathcal{L}$, $v(\phi, w) = 1$ if and only if $\phi \in \Gamma$.[22] There are thus no constraints on what sentences impossible worlds can satisfy, and there are at least as many impossible worlds as there are sets of sentences of our base language $\mathcal{L}$.[23]

As I explained in Section 1, permissive impossible-worlds models don't face the problem of logical omniscience. For instance, assume that $\langle \mathcal{M}, w\rangle \vDash B\phi$, and that ϕ logically entails ψ, i.e., that ψ is true in all the *possible* worlds in which ϕ is true. For all that we have said about the doxastic accessibility function d, it might be that $d(w)$ includes a world $w' \in \mathcal{I}$ such that $\langle \mathcal{M}, w'\rangle \nvDash \psi$, and thus that $\langle \mathcal{M}, w\rangle \nvDash B\psi$. On the other hand, the model as it stands allows any combination of beliefs and knowledge, and thus doesn't capture logical or conceptual competence. It also doesn't capture the idea that one already believes (knows) what one can easily (and correctly) compute.

Here is the main idea of my proposed supplementation of the impossible-worlds model. I propose to add the following constraints on the accessibility functions: a world w' is accessible from w if and only if w' respects what is "easily computable" from w, i.e., w' satisfies all the sentences that the agent can easily compute in w to be true, and doesn't satisfy any sentence that the agent can easily compute in w not to be true (which might not be equivalent to computing that the sentence is false). At a first pass, an agent "computes" ϕ to be true if she would answer affirmatively when asked

[22] Nolan [31, p. 542] introduces this as a comprehension principle on impossible worlds; it is also adopted by Bjerring and Skipper [8, p. 509].

[23] Note that these assumptions about the structure of impossible worlds are compatible with different accounts of the metaphysical nature of impossible worlds, e.g., accounts on which impossible (and possible) worlds are concrete entities [57], points in modal space [58], or constructions out of positive and negative facts [23, chap. 5]. For an overview of accounts of the metaphysical nature of impossible worlds, see [5, chaps. 2 and 3].

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

whether ϕ is true, given that she desires to give the correct answer. The answer could be incorrect in the doxastic case, but not in the epistemic case. In either case, an agent can compute ϕ to be true without having a proof (or what they take to be a proof) of ϕ. As we will see, we can generalize this understanding of "compute ϕ to be true" to cover non-linguistic actions as well.

The qualifier "easily" is supposed to bound the algorithmic notion of belief and knowledge as discussed in Section 2.1. In general, an agent who believes (knows) ϕ is disposed to act (capable of acting) upon ϕ—including to answer questions about ϕ—using only a small amount of resources. For instance, we wouldn't say that Ola believes that 38, 629 is prime if she would have to think for 3 weeks before answering "Yes" when asked about it. But Ola knows that 38, 624 is composite even though she has never considered the question, because it only takes her a very small amount of computational resources to answer "Yes" to the question "Is 38, 624 composite?" For simplicity, here I will only consider the resource of time. I will thus model "easily" as "in $\leq \varepsilon$ units of time," where ε is a small "threshold" natural number (assuming also that units of time can be counted in natural numbers). It is clear from the examples above that a couple of seconds count as "in $\leq \varepsilon$ units of time," while 3 weeks or 3 hours don't. But there are plausibly indeterminate cases in between. As Stalnaker notes in a similar context, this fits with the plausible view that attributions of belief and knowledge are context-sensitive:

> There is obviously a continuum here, and no very natural place to draw a line between information that is easily accessible and information that is not. I don't think this is a serious problem. Attribution of belief and knowledge are obviously highly context-dependent, and the line between what we already know and what we could come to know if we made the effort may be one thing determined somewhat arbitrarily in different ways in different situations. [48, p. 437]

I will henceforth assume that the threshold ε is a small enough unit of time in the range of a few seconds, but follow Stalnaker in allowing the value of ε to be sensitive to the context of attribution of belief or knowledge.

On this algorithmic impossible-worlds model, it will turn out that an agent believes (knows) ϕ if and only if she has an algorithm that would (correctly) output an affirmative answer if asked "Is ϕ true?" in less than or equal to ε units of time. Our worlds-based algorithmic model will thus turn out to be a combination of the non-worlds-based algorithmic models of Halpern et al. [17] and Duc [15], but where belief and knowledge are bounded (unlike Halpern et al.'s notions of belief and knowledge and Duc's notion of knowledge $K_i^{\exists}$) and not themselves time-stamped notions (unlike Duc's notions of knowledge K_i^n). The model will obviously satisfy the third desideratum on middle-ground models outlined in Section 1: an agent already believes (knows) what is easily (and correctly) computationally accessible to her. But further constraints will be needed to satisfy the other two desiderata, viz., to capture logically competent agents, and to capture conceptually competent agents.

Here, then, is the proposal in more detail. Let $\mathcal{L}$, as defined above, be the language of our model, and let an *algorithmic impossible-worlds model* be a tuple $\mathcal{M} = \langle \mathcal{W}, \mathcal{P}, d, e, v, \mathcal{A}, A \rangle$, where $\mathcal{M} = \langle \mathcal{W}, \mathcal{P}, d, e, v \rangle$ is an impossible-worlds model;

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

$\mathcal{A}$ is a set of algorithms that take as input a sentence $\phi \in \mathcal{L}$ and output either "Yes," "No," or "?";[24] and A is a *local algorithm function* that assigns each $w \in \mathcal{W}$ to a *local algorithm* in $\mathcal{A}$, which I denote "A_w." I abbreviate "A_w outputs 'Yes' given ϕ in less than or equal to n units of time" as "$A_w^{\leq n}(\phi) =$ 'Yes'," and "A_w outputs 'Yes' given ϕ in some finite unit of time" as "$A_w(\phi) =$ 'Yes'" (and do the same for the other outputs). Thus $A_w^{\leq n}(\phi) =$ "Yes" implies $A_w(\phi) =$ "Yes" and $A_w(\phi) =$ "Yes" implies that there is an $n \in \mathbb{N}$ such that $A_w^{\leq n}(\phi) =$ "Yes" (and the same holds for the other outputs).

Let me first explain what I take the local algorithms to capture. As I mentioned above, on a first-pass understanding, A_w captures the agent's dispositions in state w to answer questions about the truth-values of certain sentences. Thus $A_w(\phi) =$ "Yes" captures that in state w, the agent is such that if she were asked whether ϕ is true and wanted to give the correct answer, she would engage in a certain chain of reasoning and answer affirmatively within some finite unit of time. "No" corresponds to a negative answer, and "?" to "I don't know" or "I give up."[25] I assume that agents can have different local algorithms in different states, and that a local algorithm in a state can run different sub-algorithms given different inputs.[26] For instance, it might be that 10 years ago, Ola had no local algorithm that would check for primality if asked "Is n prime?" for any n, and that in her current state, Ola has a local algorithm that would run a trial division if asked "Is 55, 579 prime?" but directly output "No" if asked whether a number ending in 4 is prime. As in [15], I thus assume that a local algorithm captures all the steps for choosing different sub-algorithms for different inputs.

On the first-pass understanding, local algorithms capture linguistic dispositions—just as in the standard algorithmic models from Section 2.1. This understanding is simpler to work with, which is why I will adopt it in discussing the model. Importantly, however, this understanding can be generalized now that we have adopted an impossible-worlds framework. For instance, following Stalnaker's [47] definition of belief, we can take $A_w(\phi) =$ "Yes" to capture that in state w, the agent is such that, all else being equal, she would output behavior that would tend to satisfy her desires in worlds in which $[\![\phi]\!]$, together with her other beliefs, are true, within some finite unit of time (where "$[\![\phi]\!]$" stands for the proposition expressed by ϕ). "No" could then correspond to behavior that would tend to satisfy her desires in worlds in which $[\![\phi]\!]$ is not true, and "?" to behavior that would tend to satisfy her desires irrespective of whether $[\![\phi]\!]$ is true. On this generalized understanding, outputting the correct answer to the question of whether ϕ is true is only one example of behavior that would tend to satisfy the agent's desires in worlds in which $[\![\phi]\!]$, together with her other beliefs, are true; thus, $A_w(\phi) =$ "Yes" on the generalized understanding doesn't require that the agent in state w would answer affirmatively if asked about ϕ (she can misspeak, for instance) or that she has any linguistic dispositions at all. On a possible-worlds

[24] To generalize this model to credences, we could instead let the algorithms output a set of reals in [0, 1].

[25] Following Parikh [34, p. 5], we could assume that there is some large resource bound β such that after β units of time, the agent's local algorithm returns "?" (we can choose β such that β units of time is weeks, months, or even longer). We would then have to reinterpret "$A_w(\phi) =$ 'Yes'" as equivalent to "$A_w^{\leq \beta}(\phi) =$ 'Yes'."

[26] We could assume that agents only have local algorithms at possible worlds (and thus restrict A's domain to $\mathcal{P}$), but I opt for the more general construal for simplicity. See [25] for a discussion of why it makes sense to say that agents have dispositions in impossible circumstances.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

framework, the first-pass understanding of the local algorithms can't be generalized in this way, since on a possible-worlds construal of propositions, Stalnaker's definitions of belief and knowledge imply that they are closed under entailment. In the case of single-premise entailment, this is because for possible-worlds propositions A and B, if A entails B, then $A \cap B = A$, and belief distributes over intersections on Stalnaker's definition, i.e., if S believes $A \cap B$, then she believes A and she believes B. (This is because if S is disposed to satisfy desires in worlds in which $A \cap B$, together with her other beliefs, is true, she is also disposed to satisfy desires in worlds in which A, together with her other beliefs—which include $A \cap B$—is true.)[27]

On the algorithmic impossible-worlds model as I have defined it, the objects of knowledge, the objects of belief, and the inputs of local algorithms are all sentences and not propositions—just as in the standard algorithmic models from Section 2.1. But because we have adopted an impossible-worlds framework, we can now also easily modify this definition and give an alternative interpretation of the formalism: we can instead let the inputs of the algorithms in $\mathcal{A}$ be propositions, $[\![\phi]\!]$, and interpret "$K\phi$" as the formalization of the proposition that the agent knows $[\![\phi]\!]$. The reason this is possible is that unlike in the construction of propositions as sets of possible worlds, for any two sentences $\phi \neq \psi \in \mathcal{L}$, the respective propositions as sets of worlds $[\![\phi]\!] = \{w \in \mathcal{W} \mid \langle \mathcal{M}, w \rangle \vDash \phi\}$ and $[\![\psi]\!] = \{w \in \mathcal{W} \mid \langle \mathcal{M}, w \rangle \vDash \psi\}$ aren't identical. (This follows from the maximally permissive construal of impossible worlds on which there is some $w' \in \mathcal{I}$ such that for all $\alpha \in \mathcal{L}$, $v(\alpha, w') = 1$ if and only if $\alpha \in \{\phi\}$, and this $w' \in [\![\phi]\!] \setminus [\![\psi]\!]$.) I will use the original definitions here for ease of notation, but officially adopt the modified definition and interpretation.

I retain the standard definitions of satisfaction and dissatisfaction given above. The key remaining task is to define the accessibility functions d and e. We first need to contrast the epistemic case with the doxastic case. Recall that on the first-pass understanding, the general idea is that one believes ϕ just in case one would affirmatively answer the question of whether ϕ is true. Since it is standardly assumed that knowledge entails belief and is factive, a natural addition for the case of knowledge is to require that the answer would be correct. Accordingly, given an algorithmic impossible-worlds model, $\mathcal{M}$, let an algorithm $X \in \mathcal{A}$ be *veridical* about ϕ at $\langle \mathcal{M}, w \rangle$ if and only if, if $X(\phi) =$ "Yes," then $\langle \mathcal{M}, w \rangle \vDash \phi$, and if $X(\phi) =$ "No," then $\langle \mathcal{M}, w \rangle \nvDash \phi$. That is, an algorithm is veridical about ϕ at w just in case if it answers "Yes" to whether ϕ is true, then ϕ is true at w, and if it answers "No" to whether ϕ is true, then ϕ is not true at w (we relativize veridicality to a world to leave it open that different worlds can have the same local algorithm).[28] In the following, I will adopt veridicality as a minimal condition on knowledge. Plausibly, knowledge requires more than true belief. Other potential conditions could be integrated into the algorithmic framework as well. Take, for instance, the condition that knowledge is *safe* true belief, i.e., belief that is true in all nearby possible worlds.[29] Our model could incorporate this idea by requiring that a local algorithm isn't just veridical about ϕ locally but in all the

[27] See [44, p. 5; 45, p. 443] for further discussion of this feature of Stalnaker's account.

[28] On the generalized understanding, the veridicality constraint would correspond to the addition that the output behavior would satisfy the agent's desires in the world that the agent is currently in.

[29] See, e.g., [56] for a defense of safety conditions on knowledge.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

nearby worlds in which that algorithm is local.[30] Or take the *sensitivity* condition on knowledge, i.e., that the agent wouldn't believe ϕ if it were false.[31] Our model could incorporate this idea by requiring that a local algorithm doesn't output "Yes" in the nearest world(s) in which it is local and ϕ is false. In similar ways, we could require the algorithms to take information as input, or to not rely on false lemmas, and so on.[32] The algorithmic approach thus yields novel and potentially fruitful ways of formally representing conditions on knowledge, by making these conditions on agents' algorithms.[33]

We can now define the accessibility functions d and e. Following the main idea explained above, for any $w \in \mathcal{W}$:

$$\begin{aligned} d(w) := \big\{w' \in \mathcal{W} \mid \text{for all } \phi \in \mathcal{L}, \text{ if } & A_w^{\leq\varepsilon}(\phi) = \text{"Yes," then } \langle \mathcal{M}, w' \rangle \vDash \phi, \text{ and} \\ & \text{if } A_w^{\leq\varepsilon}(\phi) = \text{"No," then } \langle \mathcal{M}, w' \rangle \nvDash \phi\big\}; \\ e(w) := \big\{w' \in \mathcal{W} \mid \text{for all } \phi \in \mathcal{L}, \text{ if } & A_w^{\leq\varepsilon}(\phi) = \text{"Yes" and } A_w \text{ is veridical about } \phi \text{ at } \langle \mathcal{M}, w \rangle, \\ & \text{then } \langle \mathcal{M}, w' \rangle \vDash \phi, \text{ and} \\ & \text{if } A_w^{\leq\varepsilon}(\phi) = \text{"No" and } A_w \text{ is veridical about } \phi \text{ at } \langle \mathcal{M}, w \rangle, \\ & \text{then } \langle \mathcal{M}, w' \rangle \nvDash \phi\big\}. \end{aligned}$$

As before, ε is our *threshold* unit of time, which we assume to be fixed by the context we are modeling.

On our definition, the worlds accessible from w satisfy all the sentences that the agent can easily (and correctly) compute in w to be true, and don't satisfy any of the sentences that the agent can easily (and correctly) compute not to be true (there are no constraints on sentences for which the agent outputs "?.")[34] As desired, it follows from these definitions that for any $w \in \mathcal{P}$, (i) and (ii) are equivalent, and (iii) and (iv) are equivalent:

(i) $\langle \mathcal{M}, w \rangle \vDash B\phi$.
(ii) $A_w^{\leq\varepsilon}(\phi) =$ "Yes."
(iii) $\langle \mathcal{M}, w \rangle \vDash K\phi$.
(iv) $A_w^{\leq\varepsilon}(\phi) =$ "Yes" and A_w is veridical about ϕ at $\langle \mathcal{M}, w \rangle$.

To see that these equivalences hold, consider, first, the equivalence between (i) and (ii). Let $w \in \mathcal{P}$, and assume (i). Let $Y_w = \{\alpha \in \mathcal{L} \mid A_w^{\leq\varepsilon}(\alpha) = \text{"Yes"}\}$. By the maximally permissive construction of impossible worlds, there is some $y \in \mathcal{I}$ such that for all $\alpha \in \mathcal{L}$, $v(\alpha, y) = 1$ if and only if $\alpha \in Y_w$. Thus, for all $\alpha \in \mathcal{L}$, if $A_w^{\leq\varepsilon}(\alpha) =$ "Yes"

30 One could also have this hold for a class of propositions $\Gamma \ni \phi$ that are "similar" to ϕ, along the lines of the idea that knowledge *globally method safe*, i.e., belief produced by a method that is reliable for a class of similar propositions; see, e.g., [3] for a discussion of global method safety.

31 See, e.g., [21] for a recent defense of the sensitivity condition on knowledge.

32 One could also add a probabilistic constraint following a similar proposal by Halpern and Pucella [18] for weakening the soundness constraint discussed in Section 2.1.

33 Standard approaches to modeling the justification condition include [1, 54].

34 Halpern and Pucella [19, p. 232] in passing suggest a similar idea for conjoining the algorithmic and impossible-worlds models by letting the unique accessible world from w be the one that makes true all and only ϕ such that $A(\phi) =$ "Yes." But this construal gives up many benefits of the worlds-based epistemic framework, including the account of learning as ruling out of epistemic possibilities.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

then $\langle \mathcal{M}, y \rangle \vDash \alpha$, and if $A_w^{\leq \varepsilon}(\alpha) =$ "No" then $\langle \mathcal{M}, y \rangle \nvDash \alpha$. Thus, by the definition of d, $y \in d(w)$. Thus, by (i), $\langle \mathcal{M}, y \rangle \vDash \phi$. By the choice of y, $\phi \in Y_w$, and thus (ii) follows. For the converse, assume (ii). Let $w' \in \mathcal{W}$ be such that $w' \in d(w)$. Thus, by the definition of d, for all $\gamma \in \mathcal{L}$, if $A_w^{\leq \varepsilon}(\gamma) =$ "Yes" then $\langle \mathcal{M}, w' \rangle \vDash \gamma$. Thus, by (ii), $\langle \mathcal{M}, w' \rangle \vDash \phi$. Thus, (i) follows. A parallel argument establishes the equivalence of (iii) and (iv). It is easy to see that our definitions of d and e entail the factivity of knowledge (since e is *reflexive*, i.e., for all $w \in \mathcal{W}, w \in e(w)$) and that knowledge entails belief (since $d(w) \subseteq e(w)$ for all $w \in \mathcal{W}$).[35]

Let me now turn to the three desiderata on middle-ground models from Section 1 and explain how this algorithmic impossible-worlds models can be supplemented to satisfy the first and second desiderata, and how it already satisfies the third desideratum. Consider the first desideratum, viz., that our model should capture agents who are logically competent. At an intuitive level, there are different ways to understand what it is for an agent to be logically competent. For one, and as Duc [13, p. 6; 14, p. 637] mentions, one might want a logically competent agent to "know at least a (sufficiently) large class of logical truths." For instance, it might be that any logically competent agent will know all propositions of the form $\phi \to \phi$ (although perhaps only as long as ϕ is simple enough). Similarly, and as captured by Duc's (**NEC**^{A}), one might want a logically competent agent to be capable of eventually giving the right answer to the question of whether some sentence is a tautology in some basic logical formal system, if they have enough time to think about it. Our algorithmic framework can capture this first intuitive sense of logical competence by putting both time-sensitive and non-time-sensitive constraints on logically competent agents' local algorithms. Let F be some formal system such that all the tautologies of F are true in all possible worlds $w \in \mathcal{P}$ (for instance, F might be propositional logic), and let the *basic* tautologies of F be some set of tautologies we deem to be required for logical competence (what counts as "basic" in this sense might be context-sensitive). We can then require that for any world $w \in \mathcal{P}$ and sentence $\phi \in \mathcal{L}$:

(a) If ϕ is a tautology of F, then $A_w(\phi) =$ "Yes" and A_w is veridical about ϕ at $\langle \mathcal{M}, w \rangle$.
(b) If ϕ is a basic tautology of F, then $A_w^{\leq \varepsilon}(\phi) =$ "Yes" and A_w is veridical about ϕ at $\langle \mathcal{M}, w \rangle$.

Given the equivalence of (iii) and (iv), (b) entails that logically competent agents know all the basic tautologies.

As we saw in Section 2.1, we can also follow Duc [15] and introduce new formulas of the form "$K^{\exists}\phi$" in our object language $\mathcal{L}$ to formalize propositions such as "the agent would correctly answer 'Yes' when asked whether ϕ is true in finite time" or "the agent is in a position to know ϕ" (assuming that after having correctly computed that

[35] One possibly counterintuitive consequence of our definitions of e and d is that for any $\Gamma \subseteq \mathcal{L}$, if for each $\gamma \in \Gamma$ there is an accessible world that satisfies γ, then there is an accessible world that satisfies all of Γ. This means, for instance, that if an agent doesn't rule out ϕ and doesn't rule out $\neg\phi$, then she also doesn't rule out that ϕ and $\neg\phi$ both obtain. This consequence isn't clearly problematic, since such an agent can still rule out $(\phi \wedge \neg\phi)$ and know $\neg(\phi \wedge \neg\phi)$. In any case, one could avoid this consequence by moving to a language that can express the claim that ϕ and ψ both obtain (e.g., with a formula of the form "$\phi \sqcap \psi$"), thereby leaving open that an algorithm could output "?" to ϕ, "?" to $\neg\phi$, but "No" to "$\phi \sqcap \neg\phi$." (Kocurek [28] offers a language that has the expressive power to define such a connective.)

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

ϕ is true, an agent knows ϕ). By defining the relevant kinds of accessibility relations, we can then just as above get the equivalence between (v) and (vi) for all $w \in \mathcal{P}$:

(v) $\langle \mathcal{M}, w \rangle \vDash K^{\exists}\phi$.
(vi) $A_w(\phi) =$ "Yes" and A_w is veridical about ϕ at $\langle \mathcal{M}, w \rangle$.

We can thus in the object language capture that for any tautology, ϕ, the logically competent agent is "in a position to know ϕ" in that she would give the right answer to whether ϕ is true within a finite unit of time. Given our purposes here, however, putting constraints directly on the local algorithms suffices to constrain our model to capture logically competent agents.

On a second intuitive understanding, logical competence is a *conditional* achievement. This is captured by the statements that, for instance, a logically competent agent "can draw sufficiently many conclusions from their knowledge" [13, p. 6], "does not miss out on any *trivial* logical consequences of what she believes" [8, pp. 502f.], or that "rational agents seemingly know the trivial consequences of what they know" [24, p. 1152]. On one way of understanding them, these statements suggest a closure principle of the form: if the agent believes (knows) certain propositions Φ, then she also believes (knows) the "trivial" consequences of Φ. However, as Bjerring and Skipper [8, pp. 506–508] point out, a "collapse argument" shows that such closure principles entail or "collapse into" Full Logical Omniscience: Assume that ϕ is a *trivial* consequence of Φ if ϕ is derivable from Φ in one application of a standard rule of inference. Since any logical consequence of Φ is derivable from Φ via a chain of trivial consequences, if one fails to know some logical consequence of what one knows, then one must also fail to know some trivial consequence of what one knows.[36] The challenge for capturing the second intuitive sense of logical competence, then, is to formulate some conditional constraint(s) that doesn't (don't) face a collapse argument; let me call this "the collapse challenge." In Section 3, I will outline how Bjerring and Skipper [8] propose to meet the collapse challenge. In our algorithmic impossible-worlds approach, we can meet the collapse challenge by adopting (some of) the following conditional constraints on logically competent agents' local algorithms. For simplicity, I only formulate these constraints for the primitive logical expressions of our base language $\mathcal{L}$, viz. "$\wedge$" and "$\neg$," but similar constraints can be formulated for other logical constants. I also omit the veridicality condition, but similar constraints that include it can be formulated.[37]

Consider, first, the following non-time-sensitive constraints. For any $w \in \mathcal{P}$ and $\phi, \psi \in \mathcal{L}$:

(c) If $A_w(\phi) =$ "Yes" and $A_w(\psi) =$ "Yes," then $A_w((\phi \wedge \psi)) =$ "Yes."
(d) If $A_w((\phi \wedge \psi)) =$ "Yes," then $A_w(\phi) =$ "Yes" and $A_w(\psi) =$ "Yes."
(e) $A_w(\neg\phi) =$ "Yes" if and only if $A_w(\phi) =$ "No."
(f) $A_w(\neg\phi) =$ "No" if and only if $A_w(\phi) =$ "Yes."

In outline, constraints (c)–(f) capture the intuitive idea that a logically competent agent should respect the introduction and elimination rules for the logical connectives, even though it might take her a long time to do so. For instance, constraint (c) states that if a logically competent agent would eventually assent to ϕ and to ψ, then she would

[36] This collapse argument is also raised and discussed earlier in [6; 23, chap. 6; 24].
[37] I also assume that the logical connectives have their classical meanings, but alternative constraints can be formulated for non-classical background logics.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

also eventually assent to their conjunction. Similarly, constraint (d) states that if a logically competent agent would eventually assent to a conjunction, then she would also eventually assent to its conjuncts. Constraints (e) and (f) capture that a logically competent agent computes something to be not true just in case she computes it to be false (i.e., computes its negation to be true), and vice versa. Constraints (c)–(f), together with analogous constraints we could formulate for other logical connectives and with veridicality, thus entail that if a logically competent agent believes (knows) certain propositions Φ, and ϕ can be derived from Φ in one application of a standard rule of inference, then the agent would also eventually, though perhaps not immediately, (correctly) compute that ϕ is true. Constraints (c)–(f) thus closely capture the second intuitive understanding of logical competence.

A useful consequence of (e) and (f) is that for any $w \in \mathcal{P}$, if $\langle \mathcal{M}, w \rangle \vDash B\phi$, then $\langle \mathcal{M}, w \rangle \vDash \neg B \neg \phi$ and if $\langle \mathcal{M}, w \rangle \vDash B\neg\phi$, then $\langle \mathcal{M}, w \rangle \vDash \neg B\phi$. That is, if an agent believes ϕ, then she doesn't believe $\neg\phi$, and if she believes $\neg\phi$, then she doesn't believe ϕ. To see this, consider the former conditional (the second is equivalent to the first given the classical truth-conditions for "$\neg$"). Let $w \in \mathcal{P}$, and assume that $\langle \mathcal{M}, w \rangle \vDash B\phi$. By the equivalence of (i) and (ii), it follows that $A_w^{\leq \varepsilon}(\phi) =$ "Yes." It thus follows that $A_w(\phi) =$ "Yes." Thus, by (f), $A_w(\neg\phi) =$ "No." It thus follows that $A_w^{\leq \varepsilon}(\neg\phi) \neq$ "Yes." By the equivalence of (i) and (ii), it follows that $\langle \mathcal{M}, w \rangle \nvDash B\neg\phi$, and thus that $\langle \mathcal{M}, w \rangle \vDash \neg B\neg\phi$. Similarly, (c)–(f) together entail that if an agent believes $(\phi \wedge \psi)$, then she doesn't believe $\neg\phi$ or $\neg\psi$, and if an agent believes ϕ and ψ, she doesn't believe $\neg(\phi \wedge \psi)$. The following three closure principles thereby follow from (c)–(f): for any $\alpha, \beta, \gamma \in \mathcal{L}$ and $w \in \mathcal{P}$, if $\langle \mathcal{M}, w \rangle \vDash B\alpha$ and $\langle \mathcal{M}, w \rangle \vDash B\neg\alpha$, then $\langle \mathcal{M}, w \rangle \vDash B\beta$; if $\langle \mathcal{M}, w \rangle \vDash B(\alpha \wedge \beta)$ and $\langle \mathcal{M}, w \rangle \vDash B\neg\alpha$, then $\langle \mathcal{M}, w \rangle \vDash B\gamma$; and if $\langle \mathcal{M}, w \rangle \vDash B\alpha$, $\langle \mathcal{M}, w \rangle \vDash B\beta$, and $\langle \mathcal{M}, w \rangle \vDash B\neg(\alpha \wedge \beta)$, then $\langle \mathcal{M}, w \rangle \vDash B\gamma$. As desired, these closure principles are too weak to entail Full Logical Omniscience. Constraints (c)–(f) thus don't face a collapse argument, and we have thus met the collapse challenge for capturing the second intuitive sense of logical competence.

We can also consider the result of adding time-sensitivity to (c) and (d) as possible candidate conditional constraints on logical competence:

(g) If $A_w^{\leq i}(\phi) =$ "Yes" and $A_w^{\leq j}(\psi) =$ "Yes," then $A_w^{\leq i+j+k}((\phi \wedge \psi)) =$ "Yes" for some small $k \in \mathbb{N}$, for any $i, j \in \mathbb{N}$.

(h) If $A_w^{\leq i}((\phi \wedge \psi)) =$ "Yes," then $A_w^{\leq j}(\phi) =$ "Yes" and $A_w^{\leq k}(\psi) =$ "Yes" for some $j, k \in \mathbb{N}$ such that $j, k \leq i$ for any $i \in \mathbb{N}$.

Constraint (g) puts a limit on how long it should take a logically competent agent to say "Yes" to a conjunction if she believes each of its conjuncts. This constraint leaves open that logically competent agents can believe some conjuncts without believing their conjunction. I think this is a plausible conclusion: if the computational resources it takes for an agent to compute each of the conjuncts are close to the threshold for what counts as believing something, then it might be that the agent would just take too long to compute that a conjunction is true if asked about it, in which case we wouldn't intuitively say that the agent believes the conjunction before she performed the computation. In the case of knowledge, accepting pragmatic encroachment can further strengthen this intuition: imagine that the stakes are such that an agent doesn't count as knowing ϕ unless she can make a split-second decision about whether ϕ (say, she needs to answer a timed quiz). The agent can then know ϕ and know ψ,

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

but, assuming that it would take her longer than the allowed time to answer whether $(\phi \wedge \psi)$, fail to know $(\phi \wedge \psi)$.[38]

Constraint (h), in turn, states that a logically competent agent's belief algorithm outputs "Yes" to conjuncts in less or equal time than it outputs "Yes" to their conjunction. Typicality effects such as those involved in the conjunction fallacy might provide counterexamples to (h): an agent might be quicker to judge that a conjunction such as "Linda is a bank teller and is active in the feminist movement" is true than to judge that its conjunct "Linda is active in the feminist movement" is true, because the conjunction is more "typical."[39] One might perhaps maintain that *logically competent* agents would judge the conjunct and the conjunction to be true at least equally fast. In any case, given the equivalence of (i) and (ii), (h) entails a fourth closure principle, viz., distribution over conjunction (i.e., if a logically competent agent believes a conjunction, then she also believes its conjuncts). Although the case for distribution over conjunction isn't decisive either,[40] it is widely assumed and it doesn't by itself imply anything like Full Logical Omniscience. Constraint (h) could thus also be an additional constraint on logically competent agents, depending on how strong one wants such constraints to be.

Next, consider the second desideratum on middle-ground models, viz., that our model should capture agents who are conceptually competent with the logical connectives. In my view, the constraints we have laid out in (c)–(h) can serve just as well as constraints on conceptual competence with conjunction and negation. Let us take conjunction as our main example. On traditional *inferentialist* metasemantic theories, possessing the concept of conjunction requires inferring according to the rules of conjunction-introduction and conjunction-elimination, or, at least, inferring according to these rules when "given a chance," for instance, "when someone or something brings the conclusion to your direct attention, perhaps by querying you on the matter" [55, p. 46].[41] Our constraints (c)–(h) capture precisely such a view: they entail that an agent who believes a conjunction would assent to its conjuncts if queried about them, and that an agent who believes the conjuncts of a conjunction would eventually (though perhaps not immediately) assent to the conjunction if queried about it. Different versions of inferentialism could impose slightly different interpretations on what exactly it is to have a belief algorithm that outputs "Yes" given a proposition, but the general inferentialist picture is nicely captured by constraints along the lines of (c)–(h).

Finally, consider the third desideratum on middle-ground models, viz., that our model should entail that whatever an agent can computationally easily access is already part of what she believes or knows. Our algorithmic impossible-worlds model satisfies this desideratum because it entails the equivalences between (i) and (ii) and between (iii) and (iv): on our model, an agent believes (knows) ϕ just in case she can easily (and correctly) compute ϕ to be true. In light of our discussion of the collapse challenge above, it is worth noting that this third desideratum doesn't require a closure principle of the form: agents believe (know) whatever is easily computationally accessible *from* what they believe (know) (as opposed to whatever is easily computationally accessible

38 For discussion of pragmatic encroachment, see, e.g., [26].
39 Tversky and Kahneman [53] provide this as an example of the conjunction fallacy.
40 For discussion of the case, for distribution over conjunction, see [56, pp. 276–283].
41 See also [10] for an outline and discussion of inferentialism.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

"simpliciter"). "Easily accessing ψ *from* Φ" means that we ignore any potential computational costs of accessing Φ itself: we only consider the computational costs of accessing ψ once one has already accessed Φ. But on the dispositional understanding of belief and knowledge, believing and knowing propositions Φ are compatible with the existence of computational costs of accessing Φ: as we saw in Section 1, one can believe (know) ϕ even if it takes a short computation to access and thus be able to act on ϕ. This idea is reflected in our algorithmic impossible-worlds model. But then some proposition, ψ, can be easily accessible from what one knows, Φ, while not being easily accessible simpliciter: the computational costs can add up, or the agent might not even consider Φ in situations where she needs to act on ψ (e.g., if she is asked about whether ψ). So, one won't necessarily believe (know) every proposition that is easily accessible *from* what one believes (knows): the dispositional understanding of belief and knowledge that motivate our third desideratum is incompatible with closure principles of the type that face collapse arguments.

§3. Comparison with other candidate middle-ground models. Following ideas from Duc [13, 14] and Skipper Rasmussen [38], Bjerring and Skipper [8] have recently developed a candidate middle-ground model in our sense. Their model is a dynamic doxastic model, and it uses a notion of triviality that is connected to lengths of proofs. I will explain some disadvantages of these two features of their approach, and argue that the algorithmic impossible-worlds model can subsume its advantages.[42]

In outline, *dynamic* doxastic models are models that capture transitions between doxastic states: among other things, their language has an operator, "$\langle a \rangle$," where a is some action, and $\langle \mathcal{M}, w \rangle \vDash \langle a \rangle \phi$ just in case $\langle \mathcal{N}, w' \rangle \vDash \phi$ for some $\langle \mathcal{N}, w' \rangle$ obtained by transforming $\langle \mathcal{M}, w \rangle$ according to the rules of transformation given by action a.[43] On Bjerring and Skipper's [8] model, the relevant action is inference: they take "$\langle n \rangle \phi$" to formalize "ϕ is the case after some n steps of logical reasoning," where a *step of logical reasoning* is one application of a rule of inference of some given background logical system R [8, pp. 503, 509]. Accordingly, "$\langle n \rangle B\phi$" formalizes "the agent believes ϕ after some n steps of logical reasoning" [8, p. 509]. On their construal, a proposition, ϕ, is a *trivial* consequence of a set of propositions, Γ, just in case ϕ can be inferred from Γ within n steps of logical reasoning, where n is a small enough number [8, p. 504]. Bjerring and Skipper allow that the background system R can be sensitive to the context of belief attribution; for instance, R can be a partial or a complete proof system for classical propositional logic [8, p. 504]. Moreover, the value of n, and thus also what

[42] Solaki [41, 42] and Solaki et al. [43] develop very similar dynamic models; the arguments against the dynamic feature of Bjerring and Skipper's model apply to them as well (see fns. 44 and 47). Jago [23, 24] proposes an impossible-worlds model on which it can be indeterminate which worlds are epistemically accessible to the agent, and from which it follows that an agent can fail to know trivial consequences of what she knows, but never determinately so. His approach also uses a notion of triviality and accessibility that is connected to lengths of proofs, and so the arguments against that feature of Bjerring and Skipper's model apply to his model as well. Moreover, our discussion of constraint (g) in Section 2.2 suggests, contra Jago's model, that there are cases in which one determinately knows ϕ and ψ but determinately doesn't know $\phi \wedge \psi$. For further criticisms of vagueness-based approaches to developing a middle-ground model, see [8, pp. 516–520].

[43] See [2] for an overview of dynamic epistemic logics.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

is "trivial" in their sense, "depend[] on the cognitive resources that agents have available for logical reasoning" [8, p. 503] and hence are agent-relative. Their model results in $\langle \mathcal{M}, w \rangle \vDash \langle n \rangle B\phi$ just in case ϕ follows within n steps of logical reasoning from the truths at each world doxastically accessible from w. This means that if $\langle \mathcal{M}, w \rangle \vDash B\phi$ and ψ is a trivial consequence of ϕ, then $\langle \mathcal{M}, w \rangle \vDash \langle n \rangle B\psi$ [8, pp. 515f.]. That is, if ψ is a trivial consequence of what the agent believes, then there is some n-step piece of reasoning such that if the agent follows it, then she will come to believe ψ. This is the sense in which Bjerring and Skipper claim their model captures agents who are logically competent: on their understanding, "an agent counts as 'logically competent' just in case she has the ability to infer at least the trivial logical consequences of what she believes" [8, p. 504]. Because Bjerring and Skipper don't endorse any closure principle on beliefs, their constraint on logical competence doesn't face a collapse argument.

The first, and in my view the most important, disadvantage of Bjerring and Skipper's model is that it doesn't get our third desideratum on middle-ground models. On their model, agents don't already believe the propositions that they can easily infer or compute to be true; rather, agents can *come* to believe such propositions after performing some short piece of reasoning. But this is too weak: if an agent can easily compute ϕ to be true and is thereby able to act on the information that ϕ, then we should say that she already believes ϕ. Bjerring and Skipper themselves propose the following related test for the intuitive sense of "logical competence" that they are trying to capture:

> Suppose an agent believes p, and let q be any trivial consequence of p. We can then ask: upon being asked whether q is the case, is the agent immediately able to answer "yes"? If she is, she passes the test and counts as logically competent. For example, suppose you believe that it rains and that it rains only if the streets are wet. We can then ask: are you able to immediately answer "yes" when asked whether the streets are wet? Assuming that you are 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 hence count as logically competent in the relevant sense. [8, p. 503]

On Bjerring and Skipper's construal, an agent who is immediately able to answer "Yes" to the question of whether the streets are wet is "logically competent" in the sense that she can come to believe that the streets are wet at the end of some n-step chain of reasoning. But this isn't the correct intuitive or theoretical verdict about such cases: an agent who is able to immediately answer "Yes" to the question of whether the streets are wet should be modeled as *already* believing that the streets are wet, even before they are asked about it. Intuitively, we would clearly judge such a person as already believing that the streets are wet; in general, we clearly believe many more things than what we are currently thinking about. Theoretically, this verdict follows from the standard dispositional understanding of belief from Section 1 that generated our third desideratum. On this view, beliefs (and knowledge) are supposed to explain one's abilities to act on the basis of information, such as the ability to immediately answer "Yes" to certain questions: it is *because* I believe that the streets are wet that I am able

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

to immediately answer "Yes" when asked about it. Our algorithmic impossible-worlds model easily satisfies this third desideratum, whereas dynamic models don't.[44]

Dynamic models could provide a useful way to capture in the object language what happens after an agent performs a calculation that is obviously too long for its conclusion to count as "easily accessible" in our sense. They could thus be adapted to our algorithmic model to capture how agents can transition to new belief states. However, as I mentioned at the end of Section 2.2, the algorithmic impossible-worlds model can already be easily and minimally supplemented along the lines of Duc [15] to capture in the object language what agents would eventually do (such as answer a question) after performing a computation that takes longer than ε units of time. Our algorithmic model can thus capture an important advantage of dynamic models.

A second disadvantage of Bjerring and Skipper's model is its construal of triviality. Assume that an agent, S, believes all the propositions in Γ. The fact that there is a short chain of applications of rules of inference of R to propositions in Γ that ends with ϕ (i.e., a short *derivation* of ϕ from Γ in R) is neither sufficient nor necessary for ϕ to be intuitively "trivial" for S. It could be that ϕ has a very long derivation from Γ that is nonetheless very easy for S to output when asked about ϕ, because she can quickly retrieve it from her memory. Conversely, it could be that ϕ has a very short derivation from Γ that is very difficult for S to find, because her search algorithm is highly inefficient or the search breadth is too big. Length of derivations is generally an inadequate measure of triviality; thus, requiring logically competent agents to (be able to come to) believe all the "trivial" consequences of what they believe in this sense is also too demanding: very few (if any) people we would ordinarily think of as logically competent would be able to immediately determine the truth-value of any proposition that is shortly derivable from what they believe, given that the relevant branching factor is so big. On the algorithmic models, triviality is instead construed computationally: what is "trivial" for or "easily accessible" to an agent is what she can compute in less than ε units of time, given the algorithms that are available to her. Moreover, these algorithms could (and likely would) work very differently from a stepwise application of rules of inference of some background logical system R.[45]

This second problem is also the reason why agents who are logically competent in Bjerring and Skipper's formal construal neither clearly fit their own intuitive

[44] On the dynamic models of Solaki [41, 42] and Solaki et al. [43], at each state, an agent has both a set of rules that are "available" for her to use and a cognitive capacity. On their semantics, the dynamic operator "$\langle\rho\rangle\phi$" captures that ϕ is the case after an application of the rule of inference ρ that is both available and "affordable" given the agent's current resources. Similarly to Bjerring and Skipper's result, it then follows that "competent agents would come to know and believe consequences lying within affordable applications of rules" [43, p. 752], but these agents don't already know or believe such consequences. The first objection thus applies to these models as well.

[45] As Fagin et al. [16, pp. 405–407, 412] explain, the framework of *step logic* [12] of which Bjerring and Skipper's construal is a type can also be embedded in the algorithmic approach. One might worry that the step logic and algorithmic approaches are very similar after all given that each Turing Machine (and thus, plausibly, each agent's belief algorithm) corresponds to a formal system (for a proof of this correspondence, see [39, pp. 191f.]). But the formal system that corresponds to a given agent's algorithm would look very different from a standard logical system like R; in particular, the rules of inference would look nothing like standard rules such as modus ponens.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

characterization of "being logically competent" as being able to infer the trivial consequences of what one believes, nor pass their own intuitive test for logical competence. Consider, once again, the agent who believes that it rains and that it rains only if the streets are wet. On Bjerring and Skipper's construal, such an agent is logically competent just in case there is an n-step piece of reasoning such that *if* the agent follows it, *then* she believes that the streets are wet and thus answers "Yes" to the question of whether the streets are wet. But there is no guarantee that a logically competent agent in this sense will (immediately) follow this n-step piece of reasoning, or that she is even able to do so.[46] For instance, it might be that when she is asked whether the streets are wet, the agent never thinks to apply modus ponens to the relevant propositions, or that she first tries other steps of reasoning and only comes to perform modus ponens on the relevant propositions after a long and tedious search. Bjerring and Skipper's model thus doesn't seem to get the result that agents who are logically competent in their sense are able to infer or come to believe the trivial consequences of what they believe, or to immediately answer "Yes" when asked whether a trivial consequence of their beliefs is true. Their model thus also doesn't seem to satisfy our second desideratum, i.e., to capture "logically competent" agents in an intuitive sense.[47]

§4. Conclusion. I presented an algorithmic impossible-worlds model as a middle ground between models that entail logical omniscience and those that leave open complete logical incompetence. On this model, an agent believes (knows) a proposition just in case she is disposed to act (capable of acting) upon it. The model thereby captures the most standard understanding of belief as (grounding) a behavioral disposition, and of knowledge as (grounding) a capacity to act. I then proposed some constraints one can add to the algorithmic impossible-worlds model to capture agents who are logically and conceptually competent. These constraints capture that an agent is logically or conceptually competent only if her algorithms respect the introduction and elimination rules of the standard logical connectives; this doesn't mean that logically competent agents already know all the logical truths or all the logical consequences of what they know, only that they would eventually compute them, if given enough computational resources. Finally, I compared the algorithmic strategy for developing a middle-ground model to dynamic approaches and approaches based on step logic, and argued that the algorithmic impossible-worlds model has none of their disadvantages, and that it can subsume their advantages.

Acknowledgments. For very helpful feedback, I thank Paul Audi, Sharon Berry, Tom Donaldson, Juliet Floyd, Jens Kipper, Arc Kocurek, James Walsh, Jared Warren, Dan Waxman, an audience at the University of Konstanz, and two anonymous reviewers.

[46] Berto and Jago [5, p. 121] raise a very similar point but in response to an earlier draft where Bjerring and Skipper's test is formulated in terms of what the agent *will* answer, and not what she is *able* to answer.

[47] The models of Solaki [41, 42] and Solaki et al. [43] could avoid this problem at least for single applications of rules if the rules that are "available" to an agent at a state are understood as, e.g., the rules she would apply within a suitably small unit of time if prompted appropriately.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

BIBLIOGRAPHY

[1] Artemov, S. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, **7**, 1–36.

[2] Baltag, A., & Renne, B. (2018). Dynamic epistemic logic. In Zalta, E. N., editor. *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/entries/dynamic-epistemic/.

[3] Bernecker, S. (2020). Against global method safety. *Synthese*, **197**, 5101–5116.

[4] Berto, F., & Jago, M. (2018). Impossible worlds. In Zalta, E. N., editor. *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/fall2018/entries/impossible-worlds/.

[5] ———. (2019). *Impossible Worlds*. Oxford: Oxford University Press.

[6] Bjerring, J. C. (2013). Impossible worlds and logical omniscience: An impossibility result. *Synthese*, **190**(13), 2505–2524.

[7] Bjerring, J. C., & Schwarz, W. (2017). Granularity problems. *Philosophical Quarterly*, **67**(266), 22–37.

[8] Bjerring, J. C., & Skipper, M. (2019). A dynamic solution to the problem of logical omniscience. *Journal of Philosophical Logic*, **48**, 501–521.

[9] Boghossian, P. (2011). Williamson on the *a priori* and the analytic. *Philosophy and Phenomenological Research*, **82**(2), 488–497.

[10] ———. (2012). Inferentialism and the epistemology of logic: Reflections on Casalegno and Williamson. *Dialectica*, **66**(2), 221–236.

[11] Chalmers, D. (2011). The nature of epistemic space. In Egan, A., and Weatherson, B., editors. *Epistemic Modality*. Oxford: Oxford University Press, pp. 60–107.

[12] Drapkin, J., & Perlis, D. (1986). A preliminary excursion into step-logics. In Ghidini, C., Giodini, P., and van der Hoek, W., editors. *Proceedings of the SIGART International Symposium on Methodologies for Intelligent Systems, Knoxville Tennessee USA, October 22–24, 1986*. New York, NY: Association for Computing Machinery, pp. 262–269.

[13] Duc, H. N. (1995). Logical omniscience vs. logical ignorance: On a dilemma of epistemic logic. In Pinto-Ferreira, C., and Mamede, N. J., editors. *Progress in Artificial Intelligence. EPIA 1995. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), Funchal, Madeira Island, Portugal, October 3–6, 1995*. Berlin, Heidelberg: Springer, pp. 237–248.

[14] ———. (1997). Reasoning about rational, but not logically omniscient, agents. *Journal of Logic and Computation*, **7**(5), 633–648.

[15] ———. (2001). Resource-Bounded Reasoning about Knowledge. Ph.D. Thesis, Faculty of Mathematics and Informatics, University of Leipzig.

[16] Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about Knowledge*. Cambridge: MIT Press.

[17] Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1994). Algorithmic knowledge. In *Proceedings of the 5th Conference on Theoretical Aspects of Reasoning and Knowledge (TARK'94), Pacific Grove, CA, USA, 13–16 March 1996*. San Francisco, CA: Morgan Kaufmann, pp. 255–266.

[18] Halpern, J. Y., & Pucella, R. (2005). Probabilistic algorithmic knowledge. *Logical Methods in Computer Science*, **1**(3), 1–26.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

[19] ———. (2011). Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence*, **175**, 220–235.

[20] ———. (2012). Modeling adversaries in a logic for security protocol analysis. *Logical Methods in Computer Science*, **8**(1), 1–26.

[21] Ichikawa, J. J. (2017). *Contextualising Knowledge: Epistemology and Semantics*. Oxford: Oxford University Press.

[22] Jago, M. (2006). Logics for Resource-Bound Agents. Ph.D. Thesis, The University of Nottingham.

[23] ———. (2014a). *The Impossible: An Essay on Hyperintensionality*. Oxford: Oxford University Press.

[24] ———. (2014b). The problem of rational knowledge. *Erkenntnis*, **79**, 1151–1168.

[25] Jenkins, C. S., & Nolan, D. (2012). Dispositions impossible. *Noûs*, **46**(4), 732–753.

[26] Kim, B. (2017). Pragmatic encroachment in epistemology. *Philosophy Compass*, **12**(5), 163–196.

[27] Kipper, J., Kocurek, A. W., & Soysal, Z. (2022). The role of questions, circumstances, and algorithms in belief. In Degano, M., Roberts, T., Sbardolini, G., and Schouwstra, M., editors. *Proceedings of the 23rd Amsterdam Colloquium, Amsterdam, Netherlands, 19–21 December, 2022*. Amsterdam, Netherlands, pp. 181–187.

[28] Kocurek, A. W. (2021). Logic talk. *Synthese*, **199**, 13661–13688.

[29] Lewis, D. (1979). Attitudes *De Dicto* and *De se*. *Philosophical Review*, **88**, 513–543.

[30] ———. (1986). *On the Plurality of Worlds*. Malden, MA: Blackwell.

[31] Nolan, D. (1997). Impossible worlds: A modest approach. *Notre Dame Journal of Formal Logic*, **38**, 535–572.

[32] ———. (2013). Impossible worlds. *Philosophy Compass*, **8**(4), 360–372.

[33] ———. (2020). Impossibility and impossible worlds. In Bueno, I. O., and Shalkowski, S. A., editors. *The Routledge Handbook of Modality*. London: Routledge, pp. 40–48.

[34] Parikh, R. (1987). Knowledge and the problem of logical omniscience. In Ras, Z. W., and Zemankova, M., editors. *Methodologies for Intelligent Systems, Proceedings of the Second International Symposium, Charlotte, North Carolina, USA, October 14–17, 1987*. Amsterdam: North-Holland, pp. 432–439.

[35] ———. (2008). Sentences, belief and logical omniscience: Or what does deduction tell us? *Review of Symbolic Logic*, **1**(4), 87–113.

[36] Rantala, V. (1982). Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica*, **35**, 18–24.

[37] Schwitzgebel, E. (2019). Belief. In Zalta, E. N., editor. *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/fall2019/entries/belief/.

[38] Skipper Rasmussen, M. (2015). Dynamic epistemic logic and logical omniscience. *Logic and Logical Philosophy*, **24**, 377–399.

[39] Smith, P. (2020). *An Introduction to Gödel's Theorems* (second edition). Logic Matters. Available from: https://www.logicmatters.net/resources/pdfs/godelbook/GodelBookLM.pdf.

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press

[40] Solaki, A. (2017). Steps Out of Logical Omniscience. MSc Thesis, University of Amsterdam.

[41] ———. (2019). A dynamic epistemic logic for resource-bounded agents. In Sedlár, I., and Blicha, M., editors. *The Logica Yearbook, Hejnice, Czech Republic, June 18–22, 2018*. College Publications, pp. 229–254.

[42] ———. (2022). The effort of reasoning: Modelling the inference steps of boundedly rational agents. *Journal of Logic, Language and Information*, **31**, 529–553.

[43] Solaki, A., Berto, F., & Smets, S. (2021). The logic of fast and slow thinking. *Erkenntnis*, **86**, 733–762.

[44] Soysal, Z. (2022). A metalinguistic and computational approach to the problem of mathematical omniscience. *Philosophy and Phenomenological Research*, 1–20.

[45] Speaks, J. (2006). Is mental content prior to linguistic meaning? *Noûs*, **40**(3), 428–467.

[46] Stalnaker, R. (1976). Propositions. In MacKay, A., and Merrill, D. D., editors. *Issues in the Philosophy of Language*. New Haven: Yale University Press, pp. 79–91.

[47] ———. (1984). *Inquiry*. Cambridge, MA: MIT Press.

[48] ———. (1991). The problem of logical omniscience, I. *Synthese*, **89**, 425–440.

[49] ———. (1996). Impossibilities. *Philosophical Topics*, **24**(1), 193–204.

[50] ———. (1999). The problem of logical omniscience, II. In *Content and Context: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press, pp. 255–273.

[51] ———. (2019). *Knowledge and Conditionals: Essays on the Structure of Inquiry*. Oxford: Oxford University Press.

[52] ———. (2021). Fragmentation and singular propositions. In Borgoni, C., Kindermann, D., and Onofri, A., editors. *The Fragmented Mind*. Oxford: Oxford University Press, pp. 183–198.

[53] Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**(4), 293–315.

[54] vanBenthem, J., Fernéandez-Duque, D., & Pacuit, E. (2011). Dynamic logics of evidence-based beliefs. *Studia Logica*, **99**, 61–92.

[55] Warren, J. (2020). *Shadows of Syntax: Revitalizing Logical and Mathematical Conventionalism*. Oxford: Oxford University Press.

[56] Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

[57] Yagisawa, T. (1988). Beyond possible worlds. *Philosophical Studies*, **53**, 175–204.

[58] ———. (2010). *Worlds and Individuals, Possible and Otherwise*. Oxford: Oxford University Press.

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF ROCHESTER
532 LATTIMORE HALL
435 ALUMNI ROAD
ROCHESTER, NY 14627-0078
USA
E-mail: zeynep.soysal@rochester.edu

https://doi.org/10.1017/S1755020323000059 Published online by Cambridge University Press