A Dynamic Epistemic Logic for Resource-Bounded Agents

ANTHIA SOLAKI[1]

Abstract: In this paper, we present a dynamic epistemic logic suitable for resource-bounded agents. Our setting is informed by empirical evidence on deductive reasoning performance and therefore it avoids the problem of logical omniscience. In particular, we introduce actions capturing how the agent learns, forgets, and applies inference rules. Our model is a variant of Kripke models extended with impossible worlds and our updates modify its components (epistemic accessibility, rule availability, cognitive capacity) according to each action's effect. We further provide a sound and complete axiomatization, through a method connecting this semantic approach to logical omniscience with more syntactically-oriented ones. Finally, we use similar tools to model moderate introspective ability and thus avoid the unrealistic commitment to unbounded introspection.

Keywords: dynamic epistemic logic, logical omniscience, bounded rationality, resource-bounded agents, impossible worlds, rule-based agents

1 Introduction

The work of Hintikka (1962) paved the way for the formal study of propositional attitudes via possible-worlds semantics. Still, S5 modal logic, seen as the standard epistemic logic, faces *the problem of logical omniscience* (Fagin, Halpern, Moses, & Vardi, 1995): agents are modelled as reasoners with unbounded inferential power, always knowing anything that logically follows from what they know. For example, all tautologies are known regardless of their complexity, knowledge is closed under logical equivalence etc. However, it takes time to compute whether a complex formula is indeed a tautology; empirical evidence (e.g., on *framing effects*, Tversky & Kahneman, 1981) shows that logically equivalent statements are assessed differently by subjects, depending on their presentation. These examples suggest that cognitive effort involved in reasoning tasks needs to be accounted for.

[1] This work is funded by the Dutch Science Foundation, under the "PhDs in the Humanities" scheme (project number 322-20-018).

The S5 approach is sometimes defended due to its normative status; it models how we *ought to* reason. Yet experimental evidence suggests that mistakes in deductive reasoning are in fact systematic (Stanovich & West, 2000; Stenning & van Lambalgen, 2008). For example, people's performance in the *Wason Selection Task*[2], which essentially requires an application of modus ponens and modus tollens, is notoriously poor (Wason, 1966). Similarly, cognitive limitations affect the extent of introspection one can achieve. Apart from philosophical objections (Stroll, 1967; Williamson, 2000), the S5 axioms of *positive* and *negative* introspection ($K\phi \rightarrow KK\phi$ and $\neg K\phi \rightarrow K\neg K\phi$, respectively) are not in agreement with experimental findings (Verbrugge, 2009). We therefore seek another normative model tailored to such observations. This is why we propose modelling how a rational agent comes to know things, informed by empirical facts to ensure that "ought" actually implies "can".

To that end, we emphasize that although we are fallible and non-omniscient humans, we still are *logically competent*. Despite our failures, we still engage in bounded reasoning and introspection. As a result, we ask that agents should know those consequences that can be feasibly reached from their current epistemic state. Descriptive facts, e.g., regarding limitations of time and memory, are instrumental in determining the extent of feasibility.

The problem of logical omniscience has generally attracted much attention. One of the early suggestions (Hintikka, 1975) was to supplement possible-worlds models with *impossible* (or *non-normal*) worlds, that are not logically closed. If these are accessible to the agent, and given the common interpretation of knowledge as quantifying over accessible worlds, the closure properties for knowledge are invalidated. Still, such approaches are criticized on grounds of logical competence and explanatory power. Towards this direction, there are attempts discerning explicitly and implicitly held attitudes. In (Fagin & Halpern, 1987), agents have to be additionally *aware* of something to know it explicitly; hence their models are augmented by an *awareness function*, yielding the formulas the agent is aware of. However, aspects of the problem can be retained, logical competence may be sidestepped and it is not clear how the crucial notion of resource-boundedness can be accounted for. In (Quesada, 2009), the awareness set can be modified depending on the agent's applications of inference rules.

[2] The subject is given four cards, which have a number on one side and a color on the other. The visible sides of the cards are 3, 8, red and brown. The question of the task is: which cards must you turn over to test whether the proposition "if a card shows an even number on one side, then its opposite side is red" is true?

Bjerring and Skipper (2018) provide an impossible-worlds framework that also focuses on reasoning steps the agent takes to come to know more.

While we build on the above-mentioned attempts, we are interested in capturing how resource-boundedness affects the reasoning processes underpinning knowledge; in particular, we use Dynamic Epistemic Logic (DEL) (Baltag & Renne, 2016; van Benthem, 2011; van Ditmarsch, van der Hoek, & Kooi, 2007) to keep track of the reasoning steps available to the agent, their (orderly) applications, and the cognitive effort required. This attempt is first presented in Sect. 2, dealing with deductive reasoning alone, thereby proposing a way out of logical omniscience. In Sect. 3, we give a sound and complete axiomatization, via a method that further allows for comparative remarks between syntactic and semantic approaches to the problem. This is followed by an extension of this framework towards a balanced view to introspection, given in Sect. 4.[3]

2 Resource-bounded deductive reasoning

2.1 Syntax

To begin with, we need a logical language where the rules of deductive reasoning are explicitly introduced. This is why we define:

Definition 1 (Inference rule) *Given $\phi_1, \dots, \phi_n, \psi \in \mathcal{L}_P$, where $\mathcal{L}_P$ is the propositional language based on a set of atoms Φ, an inference rule ρ is a formula of the form $\{\phi_1, \dots, \phi_n\} \rightsquigarrow \psi$.*

We denote the set of premises and the conclusion of ρ with $pr(\rho)$ and $con(\rho)$, while $\mathcal{L}_R$ denotes the set of all inference rules. However, since we focus on an agent's *knowledge*, we are interested in *truth-preserving* rules.

Definition 2 (Translation) *The translation of a rule ρ is given by $tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \to con(\rho)$.*

Then the definition of our framework's logical language is given by:

Definition 3 (Language) *First, we fix a set of constants $T := \{c_\rho \mid \rho \in \mathcal{L}_R\} \cup \{cp\}$. Then the language $\mathcal{L}$ is built as follows:*

$$\phi ::= z_1 s_1 + \ldots + z_n s_n \geq z \mid p \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid A\rho \mid [+\rho]\phi \mid [-\rho]\phi \mid \langle\rho\rangle\phi$$

where $z_1, \dots, z_n \in \mathbb{Z}$, $z \in \mathbb{Z}^r$, $s_1, \dots, s_n \in T$, $p \in \Phi$, and $\rho \in \mathcal{L}_R$

[3] This paper is part of a recently initiated line of work, using DEL to model reasoning processes and resource-bounded agents (Berto, Smets, & Solaki, 2018; Smets & Solaki, 2018).

So the language is an extension of that of standard epistemic logic with:

- Numerical inequalities introduced to deal with cognitive effort (e.g., of the form $s_1 \geq s_2$). As we will see, this is possible because the constants of T essentially express the cognitive costs of inference rules and the agent's cognitive capacity.[4]

- A, an operator introduced to capture the agent's availability of inference rules. Specifically, $A\rho$ is to say that ρ is available to the agent, who can therefore apply it.

- Dynamic operators of the form $[+\rho]$ (resp. $[-\rho]$), such that: $[+\rho]\phi$ (resp. $[-\rho]$) says "after the agent learns (resp. forgets) ρ, ϕ is true".

- Dynamic operators of the form $\langle\rho\rangle$, such that: $\langle\rho\rangle\phi$ stands for "after applying ρ, ϕ is true".

2.2 Semantics: defining models

Our semantic model makes use of *impossible worlds*. Here, we abide by the so-called *American stance* (Berto, 2013): impossible worlds are not closed under *any* notion of logical consequence. Still, we want to build a model respecting the *minimal rationality* of agents. According to Cherniak (1986) we need a "theory of feasible inferences" where the difficulty of deductive reasoning is responsible for the agent performing *some*, but not all appropriate inferences, so in fact, we need a "well-ordering of inferences" in terms of difficulty. It is natural to connect this with the consumption of cognitive resources and use it to determine where the cutoff of an inferential chain lies. To start with, we rule out what is an obvious case of inconsistency for any logically competent agent: explicit contradictions. As the cognitive load increases while deductive reasoning evolves, we need to keep track of the cognitive costs of rules, with respect to each resource, having determined beforehand the resources (e.g., memory, attention, time etc.) considered. This is so because not all inference rules require equal cognitive effort, as indicated by experimental evidence (Johnson-Laird, Byrne, & Schaeken, 1992; Rips, 1994; Stenning & van Lambalgen, 2008). The cognitive effort will be captured by a (partial) function $c : \mathcal{L}_R \to \mathbb{N}^r$, where r is the number of resources considered. This function assigns a cost to each (sound) inference rule w.r.t. each resource. In addition, we will introduce *cognitive capacity*

[4] The choice of r, appearing in the definition of inequalities, will be made clear shortly after.

in our model, a component expressing what the agent can afford w.r.t. each resource, meant to be decreased following each rule application.

Definition 4 (Semantic model) *Given a set of r-many resources Res, a* model *is a tuple $M = \langle W^P, W^I, f, V, R, cp \rangle$ where:*

- *W^P, W^I are non-empty sets of possible and impossible worlds respectively. Let $W := W^P \cup W^I$.*
- *$f : W \to \mathcal{P}(W)$ is a function mapping each world to its set of epistemically accessible worlds.*
- *$V : W \to \mathcal{P}(\mathcal{L})$ is a function mapping each world to a set of formulas. In possible worlds, the function intuitively assigns the set of atomic formulas true at the world. In impossible worlds, the function assigns all formulas true at the world.*[5]
- *$R : W \to \mathcal{P}(\mathcal{L}_R)$ is a function yielding the rules the agent has available (i.e., has acknowledged as truth-preserving) at each world.*
- *cp denotes the agent's cognitive capacity, i.e., $cp \in \mathbb{Z}^r$, intuitively standing for what the agent can afford w.r.t. each resource.*

In accordance to the remarks made above, we ask that: [6]
Minimal Consistency (MC): $\{\phi, \neg\phi\} \not\subseteq V(w)$ for all $w \in W^I, \phi \in \mathcal{L}$
Soundness of rules (SoR): for $w \in W^P$, if $\rho \in R(w)$ then $M, w \models tr(\rho)$

2.3 Logical dynamics: learning, forgetting and applying a rule

The language contains operators for actions capable of changing the rules that are available to the agent as well as her epistemic state following a certain rule-application. The semantic effect of these actions is, as usual in DEL, captured via *model transformations*. If a formula is of the form $[]\phi$ with $[]$ such an operator, then it is evaluated in a model by examining what the truth value of ϕ is at the transformed model.

The model transformation due to learning (resp. forgetting) ρ is obtained by suitably expanding (resp. restricting) the relevant model component.

[5] For simplicity, we view the valuation function as $V := V_p \cup V_i$, where the functions V_p and V_i that take care of possible and impossible worlds are injective.

[6] Assuming that propositional formulas are evaluated as usual at possible worlds.

Definition 5 (Model transformation by learning a rule) *Given a model M, its transformation by learning a rule ρ is a model $M^{+\rho}$ with*

$$R^{+\rho}(w) = \begin{cases} R(w) \cup \{\rho\} & \textit{if } \rho \textit{ is sound} \\ R(w) & \textit{otherwise} \end{cases}$$

for all $w \in W^P$. Everything else remains as in the original model.

Definition 6 (Model transformation by forgetting a rule) *Given a model M, its transformation by forgetting a rule ρ is a model $M^{-\rho}$ with $R^{-\rho}(w) = R(w) \setminus \{\rho\}$, for all $w \in W^P$. Everything else remains as in the original.*

To capture the change induced by applications of inference rules, we have to encode them on the structure of our models. The effect of applying a rule is an expansion of the agent's information. We first introduce the notation $V^*(w)$ to restrict $V(w)$ to the propositional formulas satisfied at world w. We then impose the following condition to ensure that there are worlds capable of representing such expansions:

Succession: For every $w \in W$, if: (a) $pr(\rho) \subseteq V^*(w)$, (b) $\neg con(\rho) \notin V^*(w)$, and (c) $con(\rho) \neq \neg\phi$ for all $\phi \in V^*(w)$, then there is some $u \in W$ such that $V^*(u) = V^*(w) \cup \{con(\rho)\}$. We call u a ρ-expansion.

Next, we define the *ρ-radius*, in order to represent how ρ triggers an informational change, to the extent that *Minimal Consistency* is respected.

Definition 7 (ρ-radius) *We define the ρ-radius of a world w as follows:*

$$w^\rho := \begin{cases} \{w\}, \textit{ if } pr(\rho) \not\subseteq V^*(w) \\ \emptyset, \textit{ if } pr(\rho) \subseteq V^*(w) \textit{ and} \\ (\neg con(\rho) \in V^*(w) \textit{ or } con(\rho) = \neg\phi \textit{ for some } \phi \in V^*(w)) \\ \{u \mid u \textit{ is a } \rho\textit{-expansion of } w\}, \textit{ if } pr(\rho) \subseteq V^*(w) \textit{ and} \\ \neg con(\rho) \notin V^*(w) \textit{ and } con(\rho) \neq \neg\phi \textit{ for all } \phi \in V^*(w) \end{cases}$$

Notice that w's ρ-radius amounts to $\{w\}$ for $w \in W^P$, due to the closure of possible worlds, while the radius of an impossible world may contain another impossible world. The radius is instrumental in modifying epistemic accessibility in the transformed model, after a rule-application.

Definition 8 (Model transformation by application of a rule) *Take $M = \langle W^P, W^I, f, V, R, cp\rangle$ and $w \in W^P$. The transformation of the pointed model (M, w) by an application of ρ is the pointed model (M^ρ, w) with*

- $W^\rho = W$, $V^\rho = V$, $R^\rho = R$, *and* $cp^\rho = cp - c(\rho)$.
- $f^\rho = g$ *such that* $g(v) = \begin{cases} \bigcup\limits_{u \in f(w)} u^\rho, & \textit{for } v = w \\ f(v), & \textit{for } v \neq w \end{cases}$

That is, (M^ρ, w) is obtained by (a) replacing w's epistemically accessible worlds in M with the elements of their ρ-radii, and (b) reducing the cognitive capacity by the cost of performing the ρ-step. Notice that the properties of our models are preserved under the three operations defined.

2.4 Truth clauses

In order to give the truth clauses for our formulas, we first need to assign interpretations to the constants in T:

Definition 9 (Interpretation of terms) *Given* $M = \langle W^P, W^I, f, V, R, cp\rangle$ *parameterized by resources Res and the cognitive cost function c, the constants of T are interpreted as follows:* $cp^M = cp$ *and* $c^M_\rho = c(\rho)$.

Definition 10 (Truth clauses) *The clauses below define when a formula is* true at w in M. *For* $w \in W^I$: $M, w \models \phi$ *iff* $\phi \in V(w)$. *For* $w \in W^P$:

$$\begin{array}{lll}
M, w \models z_1 s_1 + \ldots + z_n s_n \geq z & \textit{iff} & z_1 s_1^M + \ldots + z_n s_n^M \geq z \\
M, w \models p & \textit{iff} & p \in V(w) \\
M, w \models \neg\phi & \textit{iff} & M, w \not\models \phi \\
M, w \models \phi \wedge \psi & \textit{iff} & M, w \models \phi \textit{ and } M, w \models \psi \\
M, w \models A\rho & \textit{iff} & \rho \in R(w) \\
M, w \models [+\rho]\phi & \textit{iff} & M^{+\rho}, w \models \phi \\
M, w \models [-\rho]\phi & \textit{iff} & M^{-\rho}, w \models \phi \\
M, w \models K\phi & \textit{iff} & M, w' \models \phi \textit{ for all } w' \in f(w) \\
M, w \models \langle\rho\rangle\phi & \textit{iff} & M, w \models cp \geq c_\rho,\ M, w \models A\rho \textit{ and } M^\rho, w \models \phi
\end{array}$$

Notice that our intended reading of $\geq$ is that, for example, $s_1 \geq s_2$ iff *every* i-th component of s_1 is greater or equal than the i-th component of s_2. Validity is defined with respect to possible worlds only. Given our clause for K, the presence of impossible worlds, where formulas are assigned a truth value directly rather than recursively, suffices to break the closure principles of logical omniscience. On the other hand, despite being fallible, an agent can still come to know consequences of her knowledge and gradually eliminate impossibilities she initially entertained. Consider the truth conditions for epistemic assertions like $K\phi$ prefixed by a rule ρ; they require that

(a) the rule is executable, (b) the rule is available to the agent, (c) ϕ follows from the accessible worlds via an application of ρ. Moreover, the actions of learning and forgetting rules can affect the availability, and thus account for the flexibility of actual reasoning and its possible failures. Cognitive capacity, decreasing suitably after every rule-application, determines to which extent consequences of one's knowledge can come to be in turn known. This cutoff is therefore cognitively informed and not arbitrarily fixed. Overall, this approach encompasses each rule's different contribution and effort, and explains how resources are consumed as reasoning evolves.

Example 1 Consider the following scenario: agent Alice is given 2 cards, each has a number on one side and a color (red or brown) on the other. Alice knows that if a card has an $e$ven number on one side, it has $r$ed on the other. Suppose that the 1st card has 8 on its visible side (fact denoted by e_1), and the 2nd card has brown (denoted by $\neg r_2$; r_2 stands for "the second card is $r$ed"). What should the agent derive? In seeing the even card and performing a modus ponens (MP) step (if affordable), she comes to know r_1. In seeing the brown card and performing a modus tollens (MT) step (if affordable), she comes to know $\neg e_2$. Figure 1 shows the original and the MP, MT-updated model. However, empirical evidence suggests that MT is more cognitively costly than MP (Johnson-Laird et al., 1992; Rips, 1994; Stenning & van Lambalgen, 2008). So the cost of MT exceeds the cost of MP, and it might be the case that it is so cognitively costly for Alice to apply that she cannot do so, e.g., under time pressure (consider subjects given the Wason Selection Task to complete in specific time bounds). Moreover, as many argue that MT, unlike MP, is not a primitive rule (Rips, 1994), it might be that the rule is not even available to Alice, so she needs to *learn* it and then apply it. Such scenarios exemplify how our tools can fit reasoning tasks studied in psychology of reasoning.

3 Axiomatization

3.1 Semantic and syntactic approaches

We now put forward a reduction of impossible-worlds models to Kripke models (i.e., involving solely possible worlds) augmented by syntactic functions. These functions capture the effect of impossible worlds in epistemic accessibility. In this way, we wish to combine the fallibility of real agents, for which impossible worlds stand, and the simplicity in technical manipu-

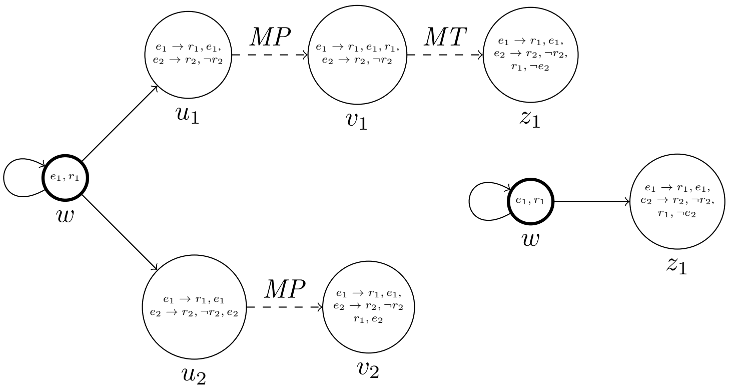

Figure 1: For worlds in W^P, unlike those in W^I, we use thicker nodes and write down only the atomic formulas satisfied there. Dashed lines indicate rule expansions. Left: the original model for Alice, who has not yet derived all consequences of her knowledge, entertaining an incomplete (u_1) and an inconsistent (u_2) world. Right: the updated model, following the applications of MP, MT.

lation of structures resembling those of Fagin and Halpern (1987). Besides, a rough division of attempts against logical omniscience is between syntactic and semantic ones. According to Fagin and Halpern (1987), a syntactic approach lacks the elegance of a semantic (impossible-worlds) one, but the latter's semantic rules do not adequately capture intuitions about knowledge and belief. To increase explanatory power and intuitiveness, we limited the arbitrariness of impossible worlds (via *Succession* and *Minimal Consistency*) and modelled logically competent agents, in that they gradually refine their epistemic state through rule-applications. In what follows, we reduce our models to syntactic structures to extract a sound and complete logic. In combination with Wansing (1990), it can be claimed that there is a "correspondence" between the two styles of attacking logical omniscience.

3.2 A common background language $\mathcal{L}_r$

In order to show that the same formulas are valid under the original and the reduced models, we introduce auxiliary operators to the *static* fragment of $\mathcal{L}$. These are such to discern the impact of possible and impossible worlds and to encode our model's structure. These operators, along with their interpretation at possible worlds, are given below:

$$\begin{array}{lll}
M,w \models L\phi & \text{iff} & M,u \models \phi \text{ for all } u \in W^P \cap f(w) \\
M,w \models I\phi & \text{iff} & M,u \models \phi \text{ for all } u \in W^I \cap f(w) \\
M,w \models \hat{I}\phi & \text{iff} & M,u \models \phi \text{ for some } u \in W^I \cap f(w) \\
M,w \models \langle RAD\rangle_\rho\phi & \text{iff} & \text{for some } u \in w^\rho\text{: } M,u \models \phi
\end{array}$$[7]

Abbreviations: (a) $[RAD]_\rho\phi := \langle RAD\rangle_\rho\top \to \langle RAD\rangle_\rho\phi$, and (b) If ϕ is of the form $\neg\psi$, for some formula ψ, then $\overline{I}\phi := \hat{I}\psi$, else $\overline{I}\phi := \bot$

3.3 The reduced model

Take a model $M = \langle W^P, W^I, f, V, R, cp\rangle$ and $V_r(w) := \{\phi \in \mathcal{L}_r \mid M,w \models \phi\}$, for $w \in W^I$. We construct $\mathfrak{M} = \langle \mathrm{W}, \mathrm{f}, \mathrm{V}, \mathrm{I}, \hat{\mathrm{I}}, \mathrm{R}, cp\rangle$ where:

- $\mathrm{W} = W^P$, $\mathrm{f}(w) = f(w) \cap \mathrm{W}$ for $w \in \mathrm{W}$, $\mathrm{V} = V|_{\mathrm{W}}$, and $\mathrm{R} = R|_{\mathrm{W}}$.
- $\mathrm{I} : \mathrm{W} \to \mathcal{P}(\mathcal{L}_r)$ such that $\mathrm{I}(w) = \bigcap_{v \in f(w) \cap W^I} V_r(v)$. Intuitively, I takes a possible world w and yields the set of those formulas that are true at all impossible worlds accessible from w.
- $\hat{\mathrm{I}} : \mathrm{W} \to \mathcal{P}(\mathcal{L}_r)$ such that $\hat{\mathrm{I}}(w) = \bigcup_{v \in f(w) \cap W^I} V_r(v)$. Intuitively, $\hat{\mathrm{I}}$ takes a possible world w and yields the set containing any formula true at some impossible world accessible from w.

The interpretation of terms in $\mathfrak{M}$ is as before, for it depends on the parameter c and the model component cp. The semantics based on $\mathfrak{M}$ is:

$$\begin{array}{lll}
\mathfrak{M},w \models z_1s_1 + \ldots + z_ns_n \geq z & \text{iff} & z_1s_1^{\mathfrak{M}} + \ldots + z_ns_n^{\mathfrak{M}} \geq z \\
\mathfrak{M},w \models p & \text{iff} & p \in \mathrm{V}(w) \\
\mathfrak{M},w \models L\phi & \text{iff} & \text{for all } u \in \mathrm{f}(w)\text{: } \mathfrak{M},u \models \phi \\
\mathfrak{M},w \models \neg\phi & \text{iff} & \mathfrak{M},w \not\models \phi \\
\mathfrak{M},w \models I\phi & \text{iff} & \phi \in \mathrm{I}(w) \\
\mathfrak{M},w \models \phi \wedge \psi & \text{iff} & \mathfrak{M},w \models \phi \text{ and } \mathfrak{M},w \models \psi \\
\mathfrak{M},w \models \hat{I}\phi & \text{iff} & \phi \in \hat{\mathrm{I}}(w) \\
\mathfrak{M},w \models A\phi & \text{iff} & \phi \in \mathrm{R}(w) \\
\mathfrak{M},w \models K\phi & \text{iff} & \mathfrak{M},w \models L\phi \text{ and } \mathfrak{M},w \models I\phi \\
\mathfrak{M},w \models \langle RAD\rangle_\rho\phi & \text{iff} & \text{for some } u \in w^\rho\text{: } \mathfrak{M},u \models \phi
\end{array}$$

Theorem 1 (Reduction) *Given a model M, construct $\mathfrak{M}$ as described above. Then $\mathfrak{M}$ is a reduction of M, i.e., for any $w \in W^P$ and formula $\phi \in \mathcal{L}_r$: $M,w \models \phi$ iff $\mathfrak{M},w \models \phi$.*

Proof. By induction on the complexity of ϕ. Recall that validity is defined with respect to possible worlds in the original model. The base case, and the steps for inequalities, $\neg$, $\wedge$ and A are straightforward. For L, I, $\hat{I}$, $\langle RAD\rangle_\rho$ we rely on the construction of the auxiliary operators and the definition of $\mathfrak{M}$. For K, the claim holds as it can be re-expressed in terms of L and I. □

3.4 Static axiomatization and reduction axioms

We first present an axiomatic system for the static part and show that it is sound and complete w.r.t. the reduced Kripke models. For the dynamic part, involving $\langle\rho\rangle$, $[+\rho]$, $[-\rho]$, we give *reduction axioms*. As usual in DEL, the static logic combined with these axioms, suffices to get a complete logic for our full framework.

Definition 11 (Axiomatization) *The static logic is axiomatized by the following axioms and the rules Modus Ponens and Necessitation (from ϕ infer $L\phi$).*

PC	All instances of classical propositional tautologies
Ineq	All instances of valid formulas about linear inequalities
K	$L(\phi \to \psi) \to (L\phi \to L\psi)$
T	$L\phi \to \phi$
MC	$\neg(I\phi \wedge I\neg\phi)$
SoR	$A\rho \to tr(\rho)$
$Succession$	$(\bigwedge_{\psi\in pr(\rho)} I\psi \wedge \neg\hat{I}\neg con(\rho) \wedge \neg\overline{I}con(\rho)) \to I\langle RAD\rangle_\rho con(\rho) \wedge$ $(I\phi \to I\langle RAD\rangle_\rho\phi)$, for $\phi \in \mathcal{L}_P$ $I\langle RAD\rangle_\rho\phi \to I\phi$, for $\phi \in \mathcal{L}_P$ and $\phi \neq con(\rho)$ $\neg\bigwedge_{\psi\in pr(\rho)} I\psi \to (I\phi \leftrightarrow I\langle RAD\rangle_\rho\phi)$, for $\phi \in \mathcal{L}_P$ $\bigwedge_{\psi\in pr(\rho)} I\psi \wedge (\hat{I}\neg con(\rho) \vee \overline{I}con(\rho)) \to I[RAD]_\rho\bot$
Red_1	$K\phi \leftrightarrow (L\phi \wedge I\phi)$
Red_2	$\langle RAD\rangle_\rho\phi \leftrightarrow \phi$
Red_3	$I[RAD]_\rho\phi \leftrightarrow (I\langle RAD\rangle_\rho\top \to I\langle RAD\rangle_\rho\phi$

Ineq, described by Fagin and Halpern (1994), is introduced to deal with inequalities.[8] MC, SoR and $Succession$ correspond to our model conditions, given how these are reflected on our language. We use T too, because this corresponds to factivity of knowledge. Red_1 reduces K in terms of L and I. Red_2 and Red_3 capture the properties of ρ-expansions. Using **M** for the class of reflexive $\mathfrak{M}$ models we show:

[8] Of course, the axioms in *Ineq* are adapted because terms are interpreted as r-tuples.

Theorem 2 (Soundness, Completeness) *The static logic is sound and complete w.r.t.* $\mathbf{M}$.

Proof. Soundness: It suffices to show that the reduction axioms are valid in this class. *Completeness*: We need to construct a suitable canonical model, corresponding to our $\mathfrak{M}$ models and their properties. This can be defined as $\mathcal{M} = \langle \mathcal{W}, \mathcal{F}, \mathcal{V}, \mathcal{I}, \hat{\mathcal{I}}, \mathcal{R}, cp \rangle$, where $\mathcal{W}, \mathcal{F}, \mathcal{V}$ are defined as usual (Blackburn, de Rijke, & Venema, 2001). Its functions are given by: $\mathcal{I}(w) = \{\phi \mid I\phi \in w\}$, $\hat{\mathcal{I}}(w) = \{\phi \mid \hat{I}\phi \in w\}$ and $\mathcal{R}(w) = \{\rho \mid A\rho \in w\}$, with $w \in \mathcal{W}$. Then we show the truth lemma (i.e., $\mathcal{M}, w \models \phi$ iff $\phi \in w$) by induction on ϕ. Completeness follows by standard modal logic results. □

Before we move to reduction axioms for our three actions, we have to express the updated terms in the language: $cp^{\rho} := cp - c_{\rho}$ and $c_{\rho}^{\rho} := c_{\rho}$.

Theorem 3 (Reduction axioms) *The following are valid in* $\mathbf{M}$*:*

$\langle \rho \rangle (z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge (z_1 s_1^{\rho} + \ldots + z_n s_n^{\rho} \geq z)$
$\langle \rho \rangle p \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge p$
$\langle \rho \rangle \neg \phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge \neg \langle \rho \rangle \phi$
$\langle \rho \rangle (\phi \wedge \psi) \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge \langle \rho \rangle \phi \wedge \langle \rho \rangle \psi$
$\langle \rho \rangle L\phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge L\phi$
$\langle \rho \rangle I\phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge I[RAD]_{\rho}\phi$
$\langle \rho \rangle K\phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge K[RAD]_{\rho}\phi$
$\langle \rho \rangle A\sigma \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge A\sigma$
$\langle \rho \rangle \hat{I}\phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge \hat{I}\langle RAD \rangle_{\rho}\phi$
$\langle \rho \rangle \langle RAD \rangle_{\rho}\phi \leftrightarrow (cp \geq c_{\rho}) \wedge A\rho \wedge \langle RAD \rangle_{\rho}\phi$

$[+\rho](z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow (z_1 s_1 + \ldots + z_n s_n \geq z)$
$[+\rho]p \leftrightarrow p$
$[+\rho]\neg\phi \leftrightarrow \neg[+\rho]\phi$
$[+\rho](\phi \wedge \psi) \leftrightarrow [+\rho]\phi \wedge [+\rho]\psi$
$[+\rho]L\phi \leftrightarrow L[+\rho]\phi$
$[+\rho]I\phi \leftrightarrow I\phi$
$[+\rho]K\phi \leftrightarrow L[+\rho]\phi \wedge I\phi$
$[+\rho]A\sigma \leftrightarrow A\sigma$, for $\sigma \neq \rho$
$[+\rho]\hat{I}\phi \leftrightarrow \hat{I}\phi$
$[+\rho]A\rho \leftrightarrow \top$, for ρ sound. $[+\rho]A\rho \leftrightarrow \bot$, otherwise
$[+\rho]\langle RAD \rangle_{\rho}\phi \leftrightarrow \langle RAD \rangle_{\rho}[+\rho]\phi$

$[-\rho](z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow (z_1 s_1 + \ldots + z_n s_n \geq z)$
$[-\rho]p \leftrightarrow p$
$[-\rho]\neg\phi \leftrightarrow \neg[-\rho]\phi$
$[-\rho](\phi \wedge \psi) \leftrightarrow [-\rho]\phi \wedge [-\rho]\psi$
$[-\rho]L\phi \leftrightarrow L[-\rho]\phi$
$[-\rho]I\phi \leftrightarrow I\phi$
$[-\rho]K\phi \leftrightarrow L[-\rho]\phi \wedge I\phi$
$[-\rho]A\sigma \leftrightarrow A\sigma$, for $\sigma \neq \rho$
$[-\rho]\hat{I}\phi \leftrightarrow \hat{I}\phi$
$[-\rho]A\rho \leftrightarrow \bot$
$[-\rho]\langle RAD\rangle_\rho\phi \leftrightarrow \langle RAD\rangle_\rho[-\rho]\phi$

Theorem 4 (Dynamic axiomatization) *The logic given by the system of Def. 11 and the reduction axioms of Th. 3 is sound and complete w.r.t.* $\mathbb{M}$.

Proof. We first check that the reduction axioms are indeed valid in $\mathbb{M}$. This, in combination with the result in Th. 2, suffices to prove the claim. □

4 Resource-bounded introspection

S5 models are reflexive, symmetric and transitive in order to capture properties of knowledge like factivity, positive and negative introspection. We have explained above that it is reasonable to impose reflexivity on our models too. However, due to the impossible worlds, symmetry and transitivity would not correspond to introspection. This is viewed as a desirable feature since avoiding unlimited introspection falls within the wider goal to model non-ideal agents. In analogy to our argumentation for resource-bounded factual reasoning, real agents are never fully introspective due to cognitive limitations therefore representation of reflective powers should be up to a certain modal depth (Ditmarsch & Labuschagne, 2007). In determining this depth, we adapt the tools used earlier for deductive reasoning.[9]

We first need to introduce introspective rules, add the respective terms and operators to the language (similar to the deductive case, e.g., $\langle\iota\rangle$ stands for "after applying introspective rule ι"), and fix cognitive costs. Then we need a model structure similar to the one provided by *Succession* to ensure that given sufficient resources, the agent can achieve higher and higher degrees of introspection. Extending V^* with the epistemic assertions satisfied

[9] Similar motivations fuel the works of Jago (2009) and Fervari and Velázquez-Quesada (2017), using actions of introspection.

at each world, we impose *Introspective Succession*: roughly, for every assertion composed by n K's there should be a successor world that validates an assertion with $(n+1)$ K's:[10]

Introspective Succession: For every $w \in W$, if: (i) $\psi \in V^*(w)$, where ψ is of the form $K^n\phi$ for some $\phi \in \mathcal{L}_P$ and $n = 0, \dots, n$, and (ii) $\neg K\psi \notin V^*(w)$, then there is some $u \in W$ such that $V^*(u) = V^*(w) \cup \{K\psi\}$.

Negative introspective succession works similarly. For simplicity, let $pr(\iota)$ denote the initial assertion and $con(\iota)$ the new one. Then, defining the introspective radius and model transformation is analogous to the deductive case. In order to get w's introspective radius (w^ι), we replace ρ with ι in Def. 7, taking cases based on the new succession condition. Then the transformation becomes:

Definition 12 (Model transformation by applying an introspective rule) *Let* $M = \langle W^P, W^I, f, V, R, cp\rangle$ *and* $w \in W^P$. *The transformation of the pointed model* (M, w) *by an application of* ι *is* (M^ι, w), *where:*

- $W^\iota = W$, $V^\iota = V$, $R^\iota = R$, *and* $cp^\rho = cp - c(\iota)$.
- $f^\iota = g$ *such that* $g(v) = \begin{cases} \bigcup\limits_{u \in f(w)} u^\iota, & \text{for } v = w \\ f(v), & \text{for } v \neq w \end{cases}$

This definition naturally leads to the truth conditions for applying ι: $M, w \models \langle\iota\rangle\phi$ iff $M, w \models cp \geq c_\iota$, $M, w \models A\iota$ and $M^\iota, w \models \phi$.

While the main idea (viewing both deduction and introspection as reasoning actions) is shared between this and earlier sections, the processes are discernible from a cognitive point of view. Both the costs and the availability of introspective rules may be treated differently, e.g., it might be that introspective rules are in principle always available and it is only the accumulated cost after each application that is responsible for limited reflection.

5 Conclusions

Overall, our dynamic logical framework overcomes logical omniscience but is also informed by empirical facts on the boundedness of real reasoners, as evinced by its formal results. The same considerations are extended to introspective abilities of agents. Meanwhile, we argue for the adequacy of this semantic approach against logical omniscience in terms of explanatory

[10] By making the model transitive, the condition is satisfied directly only for possible worlds.

power, but also show that it can be reduced to a syntactic one that enables the extraction of a sound and complete logic. Further work is especially needed on the optimal choice of cognitive parameters and on a multi-agent extension, building on the construction of inferential and introspective rules.

References

Baltag, A., & Renne, B. (2016). Dynamic epistemic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Berto, F. (2013). Impossible worlds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013 ed.). Metaphysics Research Lab, Stanford University.

Berto, F., Smets, S., & Solaki, A. (2018). The logic of fast and slow thinking. *working paper*.

Bjerring, J. C., & Skipper, M. (2018). A dynamic solution to the problem of logical omniscience. *Journal of Philosophical Logic*.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal Logic*. New York, NY, USA: Cambridge University Press.

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

Ditmarsch, H. V., & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese*, *155*(2), 191–209.

Fagin, R., & Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artif. Intell.*, *34*(1), 39–76.

Fagin, R., & Halpern, J. Y. (1994). Reasoning about knowledge and probability. *J. ACM*, *41*(2), 340–367.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning About Knowledge*. MIT press.

Fervari, R., & Velázquez-Quesada, F. R. (2017). Dynamic epistemic logics of introspection. In *International Workshop on Dynamic Logic* (pp. 82–97).

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

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic*, *4*(4), 475–484.

Jago, M. (2009). Epistemic logic for rule-based agents. *Journal of Logic, Language and Information*, *18*(1), 131–158.

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418-439.

Quesada, F. V. (2009). *Small steps in dynamics of information* (Vol. 1).

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA, USA: MIT Press.

Smets, S., & Solaki, A. (2018). The effort of reasoning: Modelling the inference steps of boundedly rational agents. In L. S. Moss, R. de Queiroz, & M. Martinez (Eds.), *Logic, Language, Information, and Computation* (pp. 307–324). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Boston, USA: MIT Press.

Stroll, A. (Ed.). (1967). *Epistemology*. New York: Harper and Rowe.

Tversky, A., & Kahneman, D. (1981, 02). The framing of decisions and the psychology of choice. *Science (New York, N.Y.)*, *211*, 453-8.

van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic Epistemic Logic* (1st ed.). Springer Publishing Company, Incorporated.

Verbrugge, R. (2009). Logic and social cognition. *Journal of Philosophical Logic*, *38*(6), 649–680.

Wansing, H. (1990). A general possible worlds framework for reasoning about knowledge and belief. *Studia Logica*, *49*(4), 523–539.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Harmondsworth: Penguin Books.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Anthia Solaki
ILLC, University of Amsterdam
The Netherlands
E-mail: `a.solaki2@uva.nl`