



# The Effort of Reasoning: Modelling the Inference Steps of Boundedly Rational Agents

Anthia Solaki<sup>1</sup> 

Accepted: 23 March 2022 / Published online: 7 May 2022  
© The Author(s) 2022

## Abstract

In this paper we design a new logical system to explicitly model the different deductive reasoning steps of a boundedly rational agent. We present an adequate system in line with experimental findings about an agent's reasoning limitations and the cognitive effort that is involved. Inspired by Dynamic Epistemic Logic, we work with dynamic operators denoting explicit applications of inference rules in our logical language. Our models are supplemented by (a) impossible worlds (not closed under logical consequence), suitably structured according to the effect of inference rules, and (b) quantitative components capturing the agent's cognitive capacity and the cognitive costs of rules with respect to certain resources (e.g. memory, time). These ingredients allow us to avoid problematic logical closure principles, while at the same time deductive reasoning is reflected in our dynamic truth clauses. We further show that our models can be reduced to awareness-like plausibility structures that validate the same formulas and we give a sound and complete axiomatization with respect to them. This approach to the agent's internal deductive reasoning is finally combined with actions of external information.

**Keywords** Logical omniscience · Bounded rationality · Inference · Dynamic epistemic logic · Impossible worlds

## 1 Introduction

We place the work in this paper against the background of investigations in AI, Game Theory and Logic on bounded rationality and the problem of logical omniscience (Fagin et al. 1995). Models of agents with unlimited inferential powers work well for certain types of distributed systems but are not sufficient to model real human reasoning and its limitations. A number of empirical studies on human reasoning

---

✉ Anthia Solaki  
a.solaki2@uva.nl

<sup>1</sup> ILLC, University of Amsterdam, Amsterdam, The Netherlands

reveal that subjects are *systematically* fallible in reasoning tasks (Stanovich and West 2000; Stenning and van Lambalgen 2008). These provide us with evidence for the fact that humans hold very nuanced propositional attitudes and performing deductive reasoning steps can only be done within a limited time-frame and at the cost of real cognitive effort. In this context a case can be made for logically competent but not infallible agents who adhere to a standard of *Minimal Rationality* (Cherniak 1986). Such an agent can make *some*, but not necessarily all, of the apparently appropriate inferences. In specifying what makes inferences (in)feasible, empirical facts pertaining to the availability of cognitive resources are crucial; for example, it is natural to take into account limitations of time and memory, when setting the standard of what the agent should achieve. As we approach this topic from the context of logic, we design a normative model, rather than a purely descriptive one.

As an illustration we consider the standard *Muddy Children Puzzle* (Fagin et al. 1995) which is based on the unrealistic assumption that children are unbounded reasoners and perfect logicians, who can perform demanding deductive steps all at once.

Suppose that  $n$  children are playing together and  $k$  of them get mud on their foreheads. Each child can see the mud on the others but not on her own forehead. First their father announces “at least one of you is muddy” and then asks over and over “does any of you know whether you are muddy?” Assuming that the kids are unbounded reasoners, the first  $k - 1$  times the father asks, everybody responds “no” but the  $k$ -th time all the muddy children answer “yes”.

We support the argument in Parikh (1987) stating that the limited capacity of humans, let alone children, can well lead to outcomes of the puzzle that are not in agreement with the standard textbook analysis. The mixture of reasoning steps a child has to take needs to be “situated” in specific bounds of time, memory, etc. As such, it is our aim in this paper to design a cognitively informed model of the dynamics of inference. To achieve this, we use tools from Dynamic Epistemic Logic (DEL) (Baltag and Renne 2016; Baltag and Smets 2008; van Benthem 2011; van Ditmarsch et al. 2007). DEL is equipped with dynamic operators, which can be used to denote applications of inference rules. We give a semantics of these operators via plausibility models (Baltag and Smets 2008). Our models are supplemented by (a) impossible worlds (not closed under logical consequence), suitably structured according to the effect of inference rules, and (b) quantitative components capturing the agent’s cognitive capacity and the costs of rules with respect to certain resources (e.g. memory, time). Note that our work, while building further on the early approaches based on impossible worlds (Hintikka 1975) to address logical omniscience, tries to overcome their main criticism of ignoring the agents’ logical competence and lacking explanatory power in terms of what really comes into play whenever we reason. In our work, deductive reasoning is reflected in the dynamic truth clauses. These include resource-sensitive ‘preconditions’ and utilize a model update mechanism that modifies the set of worlds and their plausibility, but also reduces cognitive capacity by the appropriate cost. We therefore show that an epistemic state is not expanded effortlessly, but, instead, via applications of rules, to the extent that they are cognitively affordable. We illustrate this formal setting on the above mentioned muddy children scenario for bounded rational children for the case  $k = 2$ . We further show that our models can be reduced to awareness-like plausibility

structures that validate the same formulas and a sound and complete axiomatization is given with respect to them. This paper builds further on the main ideas first presented at the WoLLIC 2018 (Smets and Solaki 2018). More specifically, it expands on Smets and Solaki (2018) by focusing also on a mixture of different types of reasoning tasks. Such tasks combine bounded reasoning with the revision of epistemic and doxastic states that occurs when the agent hears or observes new external information.

An arbitrary syntactic awareness-filter, used to discern *explicit* attitudes as in Fagin and Halpern (1987), cannot work for our purposes for it cannot be associated with logical competence. Even if ad-hoc modifications are imposed on standard awareness models, by e.g. awareness closure under subformulas, some forms of the problem are retained. A notable exception where awareness is affected by reasoning is given in Velázquez-Quesada (2011); we will pursue a similar rule-based approach in this paper. In relation to other work on tracking a fallible agent's reasoning and cognitive effort, we refer to Alechina and Logan (2009), Bjerring and Skipper (2018), and Rasmussen (2015). The first of these papers accounts for reasoning processes through, among others, inference-based state-transitions but their composition is not specified. The second includes operators for the agent's applications of inference rules, accompanied by cognitive costs, but no semantic interpretation is given. The third uses operators standing for a *number* of reasoning steps, and an impossible-worlds semantics, but it is not clear how the number of steps can be determined nor what makes reasoning halt after that. In contrast, our work aims at an elaborate unfolding of reasoning processes, that is necessary in order to provide more cognitively plausible explanations on why such processes eventually halt. In doing so, we combine the benefits of plausibility models and impossible worlds in the realistic modelling of competent but bounded reasoners. We then suggest how the technical treatment of the resulting framework can be facilitated, and embed the effect of external information in it.

The paper is structured as follows: in Sect. 2 we introduce our framework and discuss its contribution to the highlighted topics. The reduction laws (i.e. rewrite-rules) and the axiomatization are given in Sect. 3. In Sect. 4 we explain how the framework is combined with the dynamics of interaction and we finally present our conclusions and directions for further work in Sect. 5.

## 2 The Logical Framework to Model the Effort of Inference Steps

Our framework has two technical aims: (a) invalidating the closure properties of logical omniscience, and (b) elucidating the details of agents engaging in a step-wise, orderly, effortful reasoning process.

### 2.1 Syntax

Let  $\mathcal{L}_p$  denote a standard propositional language based on a set of atoms  $\Phi$ . Using this notation we first define *inference rules*:

**Definition 1** (*Inference rule*) Given  $\phi_1, \dots, \phi_n, \psi \in \mathcal{L}_p$ , an *inference rule*  $\rho$  is a formula of the form  $\{\phi_1, \dots, \phi_n\} \rightsquigarrow \psi$ , read as “whenever every formula in  $\{\phi_1, \dots, \phi_n\}$  is true,  $\psi$  is also true”.

We use  $pr(\rho)$  and  $con(\rho)$  to abbreviate, respectively, the set of premises and the conclusion of a rule  $\rho$  and  $\mathcal{L}_R$  to denote the set of all inference rules. To identify the truth-preserving rules, we define:

**Definition 2** (*Translation*) The *translation* of a rule  $\rho$  is given by the following implication in  $\mathcal{L}_p$ , i.e.  $tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \rightarrow con(\rho)$

We introduce the language  $\mathcal{L}$ , extending  $\mathcal{L}_p$  with two epistemic modalities:  $K$  for conventional knowledge, and  $\square$  for *defeasible knowledge*. As argued in Baltag and Smets (2008), it is philosophically interesting to include both attitudes in one system. While  $K$  represents an agent’s full introspective and factive attitude,  $\square$  is factive but not fully introspective. This weaker notion satisfies the S4-properties and is inspired by the defeasibility analysis of knowledge (Lehrer 2000; Stalnaker 2006), while  $K$  satisfies the S5-properties and is considered to be infallible and indefeasible. Regarding the changes of the agent’s epistemic state, induced by deductive reasoning, we introduce dynamic operators labelled by inference rules, of the form  $\langle \rho \rangle$ .

**Definition 3** (*Language  $\mathcal{L}$* ) The set of terms  $T$  is defined as  $T := \{c_\rho \mid \rho \in \mathcal{L}_R\} \cup \{cp\}$  with elements for all the cognitive costs  $c_\rho$  of inference rules  $\rho \in \mathcal{L}_R$ , and the cognitive capacity  $cp$ . Given a set of propositional atoms  $\Phi$ , the language  $\mathcal{L}$  is defined by:

$$\phi ::= p \mid z_1s_1 + \dots + z_ns_n \geq c \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid \square\phi \mid A\rho \mid \langle \rho \rangle\phi$$

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

The language comprises linear inequalities of the form  $z_1s_1 + \dots + z_ns_n \geq c$ , to deal with cognitive effort via comparisons of costs and capacity.<sup>1</sup> The modalities  $K$  and  $\square$  represent infallible and defeasible knowledge, respectively.<sup>2</sup> The operator  $A$  indicates the agent’s availability of inference rules, i.e.  $A\rho$  denotes that the agent has acknowledged rule  $\rho$  as truth-preserving (and is capable of applying it). The dynamic operators of the form  $\langle \rho \rangle$  are such that  $\langle \rho \rangle\phi$  reads “after applying the inference rule  $\rho$ ,  $\phi$  is true”. In  $\mathcal{L}$ , formulas involving  $\leq, =, -, \vee, \rightarrow$  can be defined as usual. Moreover, a formula of the form  $s_1 \geq s_2$  abbreviates  $s_1 - s_2 \geq \bar{0}$ .

### 2.2 Plausibility Models

Our semantics is based on plausibility models (Baltag and Smets 2008). In line with Spohn (1988) we use a mapping from a given set of worlds to the class of ordinals  $\Omega$  to derive the plausibility ordering. The model is augmented by impossible worlds, which

<sup>1</sup> Notice that  $c$  is an  $r$ -tuple. The choice of  $r$  is discussed in the next subsection.

<sup>2</sup> We will restrict our attention to  $K$  and  $\square$  prefixing formulas other than inequalities and  $\langle \rho \rangle$ -formulas. This is because the agent’s reflection over her own capabilities and reasoning processes, as instances of higher-order reasoning, fall outside the scope of this attempt.

need not be closed under logical consequence. However, while the agent's fallibility is not precluded—it is in fact witnessed by the inclusion of impossible worlds—it is reasoning, i.e. applications of rules, that gradually eliminates the agent's ignorance. As a starting point, we adopt a *Minimal Consistency* requirement, ruling out 'explicit contradictions' that are obvious cases of inconsistency for any (minimally) rational agent.

In order to capture the increasing cognitive load of deductive reasoning in line with empirical findings, we first introduce two parameters: (i) the agent's cognitive resources, and (ii) the cognitive cost of applying inferential rules. Regarding (i), we will use  $Res$  to denote the set of resources, which can contain *memory*, *time*, *attention*, etc. and let  $r := |Res|$  be the number of resources considered in the modelling. Regarding (ii), the cognitive effort of the agent with respect to each resource is captured by a function  $c : \mathcal{L}_R \rightarrow \mathbb{N}^r$  that assigns a *cognitive cost* to each inference rule. As the results of experiments show, not all inference rules require equal cognitive effort: Johnson-Laird et al. (1992), Rips (1994), and Stenning and van Lambalgen (2008) claim that the asymmetry in performance observed when a subject uses *Modus Ponens* and *Modus Tollens* is suggestive of an increased difficulty to apply the latter.<sup>3</sup>

Every model comes equipped with the parameters  $Res$  and  $c$ . We also introduce a *cognitive capacity* component to capture the agent's available power with respect to each resource. As resources are depleted while reasoning evolves, capacity is not constant, but it changes after each reasoning step. The choice for an agent-specific capacity that is affected by reasoning steps is in accord with connections between capacity and performance in deductive reasoning (Bara et al. 1995).

Concrete assignments of the different cognitive costs and capacity rely on empirical research. We hereby adopt a simple numerical approach to the values of resources because this seems convenient in terms of capturing the availability and cost of *time* and it is also aligned with research on *memory* (Cowan 2001; Miller 1956).<sup>4</sup>

**Definition 4 (Plausibility model)** A *plausibility model* is a tuple  $M = \langle W^P, W^I, ord, V, R, cp \rangle$  consisting of  $W^P, W^I$ , non-empty sets of possible and impossible worlds respectively.  $ord$  is a function from  $W := (W^P \cup W^I)$  to the class of ordinals  $\Omega$  assigning an ordinal to each world.  $V : W \rightarrow \mathcal{P}(\mathcal{L})$  is a valuation function mapping each world to a set of formulas.  $R : W \rightarrow \mathcal{P}(\mathcal{L}_R)$  is a function indicating the rules the agent has available (i.e. has acknowledged as truth-preserving) at each world. Cognitive capacity is denoted by  $cp$ , i.e.  $cp \in \mathbb{Z}^r$ , indicating what the agent is able to afford with regard to each resource.

<sup>3</sup> We will focus on *sound* inference rules, i.e. rules whose translation is a tautology, because (a) the agent's state is naturally built on truth-preserving inferences, and (b) it would be infeasible to (empirically) assign a cost to arbitrary arrays of premises and conclusions. This task is meaningful due to the experimental results on how humans handle rule-schemas and on how the logical complexity of the formulas involved in their instantiations relates to their difficulty (although determining the *exact* relation between the complexity of formulas and the cognitive difficulty of a rule-application depends on empirical input and is left for future work). The cost assigned to non-sound rules is thus irrelevant and will not affect our constructions.

<sup>4</sup> Numerical assignments might also pertain to the use of pupil assessment and eye-tracking as measures of attention and indicators of cognitive effort (Kahneman and Beatty 1967; Sears and Pylyshyn 2000; Xu and Chun 2009).

Regarding possible worlds, the valuation function assigns the set of atoms that are true at the world. Regarding impossible worlds, the function assigns all formulas, atomic or complex, true at the world.<sup>5</sup> The function *ord* induces a plausibility ordering, i.e. a binary relation on  $W$ : for  $w, u \in W$ :  $w \geq u$  iff  $ord(w) \geq ord(u)$ ; its intended reading is “ $w$  is no more plausible than  $u$ ”. Hence, the smaller the ordinal, the more plausible the world. The induced relation  $\geq$  is reflexive, transitive, connected and conversely well-founded.<sup>6</sup> We define  $\sim$ , representing epistemic indistinguishability:  $w \sim u$  iff  $w \geq u$  or  $u \geq w$ , i.e.  $\geq$ -comparable states are epistemically indistinguishable for the agent (Ditmarsch et al. 2015, Chapter 7).

To ensure that the rules available to the agent are truth-preserving, and assuming that propositional formulas are evaluated as usual in possible worlds, we impose *Soundness of Rules*: for every  $w \in W^P$ , if  $\rho \in R(w)$  then  $M, w \models tr(\rho)$ . We also need a condition to hardwire the effect of deductive reasoning in the model. To that end, we take:

**Definition 5 (Propositional truths)** Let  $M$  be a model and  $w \in W$  a world of the model. If  $w \in W^P$ , its set of *propositional truths* is  $V^*(w) = \{\phi \in \mathcal{L}_p \mid M, w \models \phi\}$ . If  $w \in W^I$ ,  $V^*(w) = \{\phi \in \mathcal{L}_p \mid \phi \in V(w)\}$ .

Based on  $V^*$ , which is determined by  $V$ , we impose *Succession* on the model: for every  $w \in W$ , if (i)  $pr(\rho) \subseteq V^*(w)$ , (ii)  $\neg con(\rho) \notin V^*(w)$ , (iii)  $con(\rho) \neq \neg\phi$  for all  $\phi \in V^*(w)$  then there is some  $u \in W$  such that  $V^*(u) = V^*(w) \cup \{con(\rho)\}$ .

**Definition 6 ( $\rho$ -radius)** The  $\rho$ -radius of a world  $w$  is given by:<sup>7</sup>

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

The  $\rho$ -radius, inspired by Bjerring and Skipper (2018), represents how the rule  $\rho$  triggers an informational change and its element, if it exists, is called  $\rho$ -*expansion*. A rule whose premises are not true at a world does not trigger any change, this is why the only expansion is the world itself. A rule that leads to an explicit contradiction forms the empty radius as is arguably the case for minimally rational agents. If the conditions of *Succession* are met, the radius contains the new “enriched” world. Due to the injectiveness of  $V_p$  and  $V_i$ , a world’s  $\rho$ -expansion is unique. As  $\rho$ -expansions

<sup>5</sup> We will assume that worlds are valuation-wise unique, i.e. we view the valuation as  $V := V_p \cup V_i$ , where the functions  $V_p$  and  $V_i$  taking care of possible and impossible worlds are injective. This assumption is not vital but it serves the simplicity of the setting because we avoid a multiplicity of worlds unnecessary for our purposes.

<sup>6</sup> These properties, which follow from the definition of *ord*, will not force unnecessarily strong (introspective) validities for non-ideal agents because of the presence of impossible worlds.

<sup>7</sup> Note that  $=$  between formulas stands for syntactic identity. It is used due to *Minimal Consistency* and the fact that  $V^*$  is given directly by  $V$  in impossible worlds.

expand the state from which they originate, inferences are not defeated as reasoning steps are taken, hence *Succession* warrants monotonicity, to the extent that *Minimal Consistency* is respected. Note that  $w$ 's  $\rho$ -expansion amounts to itself for  $w \in W^P$  (due to the deductive closure of possible worlds) while an impossible world's  $\rho$ -expansion is another impossible world.

### 2.3 Model Transformations and Semantic Clauses

To evaluate  $(\rho)\phi$ , we have to examine the truth value of  $\phi$  in a transformed model, defined in such a way to capture the effect of applying  $\rho$ . Roughly, a pointed plausibility model  $(M', w)$  (which consists of a plausibility model and a point indicating the real world) is the  $\rho$ -update of a given pointed plausibility model, whenever the set of worlds is replaced by the worlds reachable by an application of  $\rho$  on them, while the ordering is accordingly adapted. That is, if a world  $u$  was initially entertained by the agent, but after an application of  $\rho$  does not "survive", then it is eliminated. This world must have been an impossible world and a deductive step uncovered its impossibility. Once such worlds are ruled out, the initial ordering is preserved to the extent that it is unaffected by the application of the rule. More concretely, let  $M = \langle W^P, W^I, ord, V, R, cp \rangle$  be a plausibility model and  $(M, w)$  the pointed model based on  $w$ . The updated  $M^\rho$  model is given via the following:

- Step 1 Given a rule  $\rho$ ,  $W^\rho := \bigcup_{v \in W} v^\rho$ . In words,  $W^\rho$  consists of the  $\rho$ -expansions of the worlds initially entertained by the agent. So the  $\rho$ -updated pointed model  $(M^\rho, w)$  should be such that its set of worlds is  $W^\rho$ . As observed above, any elimination of worlds is in fact an elimination affecting the set  $W^I$ .
- Step 2 We now develop the new ordering  $ord^\rho$  following the application of the inference rule. Take  $u \in W^\rho$ . This means that there is at least one  $v \in W$  such that  $\{u\} = v^\rho$ . Denote the set of such  $v$ 's by  $N$ . Then  $ord^\rho(u) = ord(z)$  for  $z \in \min(N)$ . Therefore, if a world is in  $W^\rho$ , then it takes the position of the most plausible of the worlds from which it originated.
- Step 3  $V$  and  $R$  are simply restricted to the worlds in  $W^\rho$  and  $cp^\rho := cp - c(\rho)$ . Again, for  $u, v \in W^\rho$ , we say:  $u \geq^\rho v$  iff  $ord^\rho(u) \geq ord^\rho(v)$ . It is easy to check that all the required properties are preserved.

Prior to defining the truth clauses we need to assign interpretations to the terms in  $T$ . Their intended reading is that those of the form  $c_\rho$  correspond to the cognitive costs of inference rules whereas those of the form  $cp$  correspond to the agent's cognitive capacity. This is why  $cp$  is used both as a model component and as a term of our language. The use can be understood from the context.

**Definition 7** (*Interpretation of terms*) Given a model  $M$ , the terms of  $T$  are interpreted as follows:  $cp^M = cp$  and  $c_\rho^M = c(\rho)$ .

Our intended reading of  $\geq$  is that  $s \geq t$  iff every  $i$ -th component of  $s$  is greater or equal than the  $i$ -th component of  $t$ . The semantic clause for a rule-application should reflect that the rule must be "affordable" to be executable; the agent's cognitive capacity must endure the resource consumption caused by firing the rule. The semantics is finally given by:

**Definition 8** (*Plausibility semantic clauses*) The following clauses inductively define when a formula  $\phi$  is true at  $w$  in  $M$  (notation:  $M, w \models \phi$ ) and when  $\phi$  is false at  $w$  in  $M$  (notation:  $M, w \not\models \phi$ ). For  $w \in W^I$ :  $M, w \models \phi$  iff  $\phi \in V(w)$ , and  $M, w \not\models \phi$  iff  $\neg\phi \in V(w)$ . For  $w \in W^P$ , given that the boolean cases are standard:

$M, w \models p$ iff $p \in V(w)$ , where $p \in \Phi$	$M, w \not\models \phi$ iff $M, w \models \neg\phi$
$M, w \models K\phi$ iff $M, u \models \phi$ for all $u \in W$	$M, w \models A\rho$ iff $\rho \in R(w)$
$M, w \models \Box\phi$ iff $M, u \models \phi$ for all $u \in W$ such that $w \geq u$	
$M, w \models (\rho)\phi$ iff $M, w \models (c\rho \geq c_\rho), M, w \models A\rho$ and $M^\rho, w \models \phi$	
$M, w \models z_1s_1 + \dots + z_ns_n \geq c$ iff $z_1s_1^M + \dots + z_ns_n^M \geq c$	

Validity is defined with respect to possible worlds only. The truth clause for knowledge is standard, except that it also quantifies over impossible worlds. The truth of rule-availability is determined by the corresponding model function. It is then evident that the truth conditions for epistemic assertions prefixed by a rule  $\rho$  are sensitive to the idea of resource-boundedness, unlike plain assertions. The latter require that  $\phi$  is the case throughout the quantification set, even at worlds representing inconsistent/incomplete scenarios. The former ask that the rule is affordable, available to the agent, and that  $\phi$  follows from the accessible worlds via  $\rho$ . Since the agent also entertains impossible worlds, she has to take a step in order to gradually minimize her ignorance.

### 2.4 Discussion

These constructions overcome logical omniscience, while still accounting for how we perform inferences lying within suitable applications of rules. In particular, the argument of impossible worlds suffices to invalidate the closure principles. Moreover, the truth conditions for  $\langle \ddagger \rangle \spadesuit \phi$ , where  $\langle \ddagger \rangle$  abbreviates a sequence of inference rules and  $\spadesuit$  stands for a propositional attitude such as  $K$  or  $\Box$ , demonstrate that an agent can come to know  $\phi$  via following an *affordable* and *available* reasoning track. In fact, the rule-sensitivity, the measure on cognitive capacity and the way it is updated allow us to practically witness to what extent reasoning evolves. Besides, running out of resources depends not only on the *number* but also on the *kind* and *chronology* of rules. Our approach takes these factors into account and explains how the agent exhausts her resources while reasoning.

Unlike (Bjerring and Skipper 2018; Duc 1997) we abstain from a generic notion of reasoning process and we do not presuppose the existence of an arbitrary cutoff on reasoning. Instead, we account explicitly for (a) specific rules available to the agent, (b) their individual applications, (c) their chronology, and (d) their cognitive consumption. This elaborate analysis is crucial in bridging epistemic frameworks with empirical facts for it exploits studies in psychology of reasoning that usually



study *individual* inference rules in terms of cognitive difficulty.<sup>8</sup> Furthermore, the enterprise of providing a semantics contributes to Rasmussen (2015)'s attempt, who tracks reasoning processes, but lacks a principled way to defend his selection of axioms. Constructing a semantic model that captures the change triggered by rule-applications allows for a definition of validity important in assessing the adequacy of the solution.

We will illustrate our framework on the *Muddy Children Puzzle*, highlighted in the introduction. We analyze the failure of applying a sequence of rules in the  $k = 2$  scenario, attributed to the fact that the first rule applied is so cognitively costly for a child that her available time expires before she can apply the next. It thus becomes clear why in even more complex cases (e.g. for  $k > 2$ ) human agents are likely to fail, contrary to predictions of standard logics, whereby demanding reasoning steps are performed at once and without effort. Our framework models the dynamics of inference and the resource consumption each step induces.

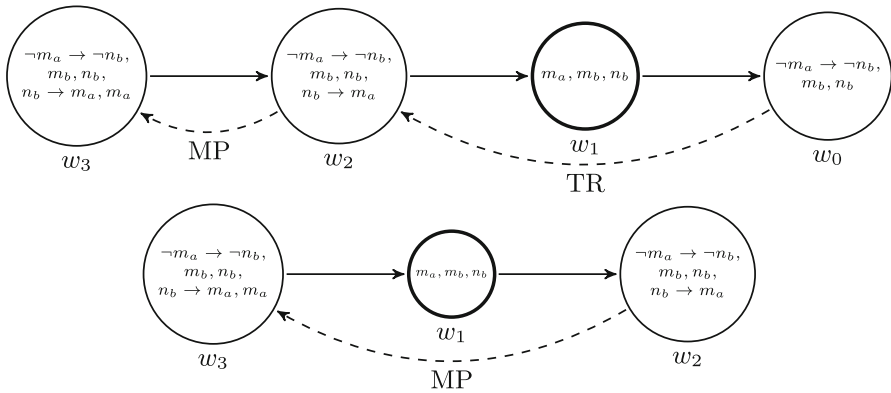
**Example 1 (Bounded muddy children)** Take  $m_a, m_b$  as the atoms for “child  $a$  (resp.  $b$ ) is muddy” and  $n_a, n_b$  for “child  $a$  (resp.  $b$ ) answers no to the first question”. Let  $M = \langle W^P, W^I, ord, V, R, cp \rangle$  be as depicted in Fig. 1. For simplicity, take two rules, transposition of the implication and modus ponens, so that  $R = \{TR, MP\}$  where  $TR = \{\neg m_a \rightarrow \neg n_b\} \rightsquigarrow n_b \rightarrow m_a$ ,  $MP = \{n_b, n_b \rightarrow m_a\} \rightsquigarrow m_a$ ,  $Res = \{time, memory\}$ ,  $c(TR) = (5, 2)$ ,  $c(MP) = (2, 2)$ ,  $cp = (5, 7)$ .

Analyzing the reasoning of child  $a$  (see Fig. 1) after the father's announcement and after child  $b$  answered “no” to the first question, we verify that  $\Box(\neg m_a \rightarrow \neg n_b)$  and  $\Box n_b$  are valid, i.e. child  $a$  initially knows that if she is not muddy, then child  $b$  should answer “yes” (as in that case only  $b$  is muddy), and that  $b$  said “no”. Following a  $TR$ -application, the world  $w_0$  is eliminated and its position is taken by its  $TR$ -expansion, i.e.  $w_2$  and  $cp^{TR} = (5, 7) - (5, 2) = (0, 5)$ . In addition,  $A(TR)$ , and  $cp \geq c_{TR}$ . Therefore  $\langle TR \rangle \Box(n_b \rightarrow m_a)$  is also valid. But now the cost of the next step is too high, i.e.  $M^{TR}, w_1 \not\models cp \geq c_{MP}$  (compare  $cp^{TR}$  and  $c(MP)$ ), so overall the formula  $\langle TR \rangle \neg \langle MP \rangle \Box m_a$  is indeed valid.

### 3 Reduction and Axiomatization

Work in Wansing (1990) shows how various models for knowledge and belief, including structures for awareness (Fagin and Halpern 1987), can be viewed as impossible-worlds models (more specifically, *Rantala models*, Rantala 1982), that validate precisely the same formulas (given a fixed background language). In the remainder, we explore the other direction and show that *our* impossible-worlds framework can be reduced to an *awareness-like* one, that only involves possible worlds. In the absence of impossible worlds, standard techniques used in axiomatizing DEL settings (via reduction axioms) can be used. This reduction is a technical contribution;

<sup>8</sup> See, for example, Cherniak (1986), Johnson-Laird et al. (1992), Rips (1994), and Schroyens and Schaeken (2003); Schroyens et al. (2001). In fact, different schools in psychology of reasoning attribute inferential asymmetries to different causes. However, the very observation that not all inferences require equal cognitive effort is common ground.



**Fig. 1** The reasoning of boundedly rational child  $a$ . Thicker borders are used for deductively closed possible worlds. In impossible worlds, we write all propositional formulas satisfied and indicate (non-trivial) expansions via dashed arrows.

the components of the reduced model lack the intuitive readings of the original framework, but allow us to prove completeness. Further, this method has the advantage of combining the benefits of impossible worlds in modelling non-ideal agents and the technical treatment facilitated by awareness-like DEL structures.

First, we show how the *static* part of the impossible-worlds setting can be transformed into one that merely involves possible worlds and captures the effect of impossible worlds via the introduction of auxiliary modalities and syntactic, *awareness-like* functions. Second, we obtain a sound and complete axiomatization for the static part through the construction of a suitable *canonical model*. Third, we give DEL-style *reduction axioms* that reduce formulas involving the dynamic rule-operators to formulas that contain no such operator. In this way, we use the completeness of the static part to get a complete axiomatization for the dynamic setting.

### 3.1 The (Static) Language for the Reduction

We first fix an appropriate language  $\mathcal{L}_r$  as the “common ground” needed to show that the reduction is successful, i.e. that the same formulas are valid under the original and the reduced models. As before, let  $\spadesuit$  stand for  $K$  or  $\square$  and take the *quantification set*  $Q_\spadesuit(w)$  to be  $W$  if  $\spadesuit = K$ , and  $Q_\spadesuit(w) := \{u \mid w \geq u\}$ , if  $\spadesuit = \square$  (to denote the set that the truth clauses for  $K$  and  $\square$  quantify over). Auxiliary operators are then introduced to the static fragment of  $\mathcal{L}$ , in order to capture (syntactically) the effect of impossible worlds in the interpretations of propositional attitudes. For  $w \in W^P$ :

- $M, w \models L_\spadesuit\phi$  iff  $M, u \models \phi$  for all  $u \in W^P \cap Q_\spadesuit(w)$
- $M, w \models I_\spadesuit\phi$  iff  $M, u \models \phi$  for all  $u \in W^I \cap Q_\spadesuit(w)$

That is,  $L_\spadesuit$  provides the standard quantification over the possible worlds while  $I_\spadesuit$  isolates the impossible worlds, for each  $\spadesuit = K, \square$ . In addition, we introduce operators to encode the model’s structure:

- $M, w \models \hat{I}_\spadesuit\phi$  iff  $M, u \models \phi$  for some  $u \in W^I \cap Q_\spadesuit(w)$

–  $M, w \models \langle RAD \rangle_\rho \phi$  iff for some  $u \in w^\rho$ :  $M, u \models \phi$

The operators  $\langle RAD \rangle_\rho$ , labelled by inference rules are such to ensure that there is some  $\phi$ -satisfying  $\rho$ -expansion. To express that *all*  $\rho$ -expansions are  $\phi$ -satisfying, we use  $[RAD]_\rho \phi := \langle RAD \rangle_\rho \top \rightarrow \langle RAD \rangle_\rho \phi$ , because once an expansion exists, it is unique. Indexed operators of this form provide information on the model’s structure; they are introduced syntactically only as temporal-style projections of connections induced by inference rules on the model. This is why their interpretation should be independent of the distinction between possible and impossible worlds. For example, for  $w \in W^I$ :  $M, w \models \langle RAD \rangle_\rho \phi$  iff for some  $u \in w^\rho$ :  $M, u \models \phi$ . We also use the following abbreviation: if  $\phi$  is of the form  $\neg\psi$ , for some formula  $\psi$ , then  $\bar{I}_\blacklozenge \phi := \hat{I}_\blacklozenge \psi$ , else  $\bar{I}_\blacklozenge \phi := \perp$ .

### 3.2 Building the Reduced Model

Towards interpreting the auxiliary operators in the reduced model, we construct *awareness-like functions*. Take  $V^+(w) := \{\phi \in \mathcal{L}_r \mid M, w \models \phi\}$  for  $w \in W^I$  and:

- $I_\blacklozenge : W^P \rightarrow \mathcal{P}(\mathcal{L}_r)$  such that  $I_\blacklozenge(w) = \bigcap_{v \in W^I \cap Q_\blacklozenge(w)} V^+(v)$ . Intuitively,  $I_\blacklozenge$  takes a possible world  $w$  and yields the set of those formulas that are true at all impossible worlds in its quantification set.
- $\hat{I}_\blacklozenge : W^P \rightarrow \mathcal{P}(\mathcal{L}_r)$  such that  $\hat{I}_\blacklozenge(w) = \bigcup_{v \in W^I \cap Q_\blacklozenge(w)} V^+(v)$ . Intuitively,  $\hat{I}_\blacklozenge$  takes a possible world  $w$  and yields the set of those formulas that are true at some impossible world in its quantification set.

The function *ord* captures plausibility and the “world-swapping” effect of rule-applications. Since the latter will be treated via reduction axioms, we provide a reduced model equipped with a standard binary plausibility relation (to serve as an *awareness-like plausibility structure* (ALPS), with respect to which the static logic will be developed). Given the original model  $M = \langle W^P, W^I, ord, V, R, cp \rangle$ , our reduced model is the tuple  $\mathbf{M} = \langle W, \geq, \sim, V, R, cp, I_\blacklozenge, \hat{I}_\blacklozenge \rangle$  where:

$$\begin{array}{ll}
 W = W^P & V(w) = V(w) \text{ for } w \in W \\
 u \geq w \text{ iff } ord(u) \geq ord(w), \text{ for } w, u \in W & R(w) = R(w) \text{ for } w \in W \\
 u \sim w \text{ iff } u \geq w \text{ or } w \geq u, \text{ for } w, u \in W & I_\blacklozenge, \hat{I}_\blacklozenge \text{ as explained before}
 \end{array}$$

The clauses based on the reduced model are such that the auxiliary operators are interpreted via the awareness-like functions. They are presented in detail in the Appendix, along with the proof that the reduction is indeed successful:

**Theorem 1** (Reduction) *Given a model  $M$ , let  $\mathbf{M}$  be its (candidate) reduced model. Then  $\mathbf{M}$  is indeed a reduction of  $M$ , i.e. for any  $w \in W^P$  and formula  $\phi \in \mathcal{L}_r$ :  $M, w \models \phi$  iff  $\mathbf{M}, w \models \phi$ .*

**Table 1** The static axioms

AXIOMS			
<i>PC</i>	All instances of classical propositional tautologies		
<i>Ineq</i>	All instances of valid formulas about linear inequalities		
$L_K$	The S5 axioms for $L_K$	<i>Soundness of Rules</i>	$A\rho \rightarrow tr(\rho)$
$L_{\square}$	The S4 axioms for $L_{\square}$	<i>Minimal Consistency</i>	$\neg(I_{\square}\phi \wedge I_{\square}\neg\phi)$
<i>Succession<sub>1</sub></i>	$(\bigwedge_{\psi \in pr(\rho)} I_{\blacklozenge}\psi \wedge \neg\hat{I}_{\blacklozenge}\neg con(\rho) \wedge \neg\bar{I}_{\blacklozenge}con(\rho)) \rightarrow I_{\blacklozenge}\langle RAD \rangle_{\rho} con(\rho) \wedge (I_{\blacklozenge}\phi \rightarrow I_{\blacklozenge}\langle RAD \rangle_{\rho}\phi)$ , for $\phi \in \mathcal{L}_p$		
<i>Succession<sub>2</sub></i>	$I_{\blacklozenge}\langle RAD \rangle_{\rho}\phi \rightarrow I_{\blacklozenge}\phi$ , for $\phi \in \mathcal{L}_p$ and $\phi \neq con(\rho)$		
<i>Succession<sub>3</sub></i>	$\neg\bigwedge_{\psi \in pr(\rho)} I_{\blacklozenge}\psi \rightarrow (I_{\blacklozenge}\phi \leftrightarrow I_{\blacklozenge}\langle RAD \rangle_{\rho}\phi)$ , for $\phi \in \mathcal{L}_p$		
<i>Succession<sub>4</sub></i>	$\bigwedge_{\psi \in pr(\rho)} I_{\blacklozenge}\psi \wedge (\hat{I}_{\blacklozenge}\neg con(\rho) \vee \bar{I}_{\blacklozenge}con(\rho)) \rightarrow I_{\blacklozenge}[RAD]_{\rho}\perp$		
<i>Local Connectedness</i>	$L_K(\phi \vee L_{\square}\psi) \wedge L_K(\psi \vee L_{\square}\phi) \rightarrow (L_K\phi \vee L_K\psi)$		
<i>Red<math>_{\blacklozenge}</math></i>	$\blacklozenge\phi \leftrightarrow (L_{\blacklozenge}\phi \wedge I_{\blacklozenge}\phi)$	<i>Indefeasibility</i>	$L_K\phi \rightarrow L_{\square}\phi$
<i>Radius<sub>1</sub></i>	$\langle RAD \rangle_{\rho}\phi \leftrightarrow \phi$		$I_K\phi \rightarrow I_{\square}\phi$
<i>Radius<sub>2</sub></i>	$I_{\blacklozenge}[RAD]_{\rho}\phi \rightarrow (I_{\blacklozenge}\langle RAD \rangle_{\rho}\top \rightarrow I_{\blacklozenge}\langle RAD \rangle_{\rho}\phi)$		

### 3.3 Axiomatization

We have reduced plausibility models to ALPSs. We now develop the (static) logic  $\Lambda$ , showing that it is sound and complete with respect to them.

**Definition 9** (*Axiomatization of  $\Lambda$* )  $\Lambda$  is axiomatized by Table 1 and the rules *Modus Ponens*, *Necessitation<sub>K</sub>* (from  $\phi$ , infer  $L_K\phi$ ) and *Necessitation<sub>□</sub>* (from  $\phi$ , infer  $L_{\square}\phi$ ).

*Ineq*, described in Fagin and Halpern (1994), is introduced to accommodate the linear inequalities.<sup>9</sup> The S5 axioms for  $L_K$  and S4 axioms for  $L_{\square}$  mimic the behaviour of  $K$  and  $\square$  in the usual plausibility models: these operators quantify over possible worlds only. The (clusters of) axioms about *Soundness of Rules*, *Minimal Consistency* and *Succession* take care of the respective model conditions (to the extent that these affect our language, given its expressiveness). The same holds for *Indefeasibility* and *Local Connectedness*, which also mimic their usual plausibility counterparts. To capture the behaviour of radius, we also introduce the *Radius* axioms. Finally, *Red $_{\blacklozenge}$*  expresses  $K$  and  $\square$  in terms of the corresponding auxiliary operators. We now move to the following theorems; the proofs can be found in the Appendix.

**Theorem 2** (Soundness)  $\Lambda$  is sound with respect to ALPSs.

**Theorem 3** (Completeness)  $\Lambda$  is complete with respect to non-standard<sup>10</sup> ALPSs.

Given the static logic, it suffices to reduce formulas involving  $\langle \rho \rangle$  in order to get a dynamic axiomatization. It is useful to abbreviate updated terms in our language as follows:  $cp^{\rho} := cp - c_{\rho}$  and  $c_{\rho}^{\rho} := c_{\rho}$ .

<sup>9</sup> *Ineq* is of course slightly adapted as terms are interpreted as  $r$ -tuples. This makes no difference for the axioms in Fagin and Halpern (1994), with the exception of *dichotomy* which is not needed given our reading of inequality.

<sup>10</sup> More on this terminology can be found in the Appendix.

**Theorem 4** (Reducing  $\langle \rho \rangle$ ) *The following are valid in the class of our models:*

$$\langle \rho \rangle (z_1 s_1 + \dots + z_n s_n \geq c) \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge (z_1 s_1^p + \dots + z_n s_n^p \geq c)$$

$$\langle \rho \rangle \langle \text{RAD} \rangle_\rho \phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \langle \rho \rangle \phi$$

$\langle \rho \rangle p \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge p$ $\langle \rho \rangle (\phi \wedge \psi) \leftrightarrow \langle \rho \rangle \phi \wedge \langle \rho \rangle \psi$ $\langle \rho \rangle I_\blacklozenge \phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge I_\blacklozenge [\text{RAD}]_\rho \phi$ $\langle \rho \rangle A\sigma \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge A\sigma$	$\langle \rho \rangle \neg \phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \neg \langle \rho \rangle \phi$ $\langle \rho \rangle L_\blacklozenge \phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge L_\blacklozenge \langle \rho \rangle \phi$ $\langle \rho \rangle \blacklozenge \phi \leftrightarrow \langle \rho \rangle L_\blacklozenge \phi \wedge \langle \rho \rangle I_\blacklozenge \phi$ $\langle \rho \rangle \hat{I}_\blacklozenge \phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_\blacklozenge \langle \text{RAD} \rangle_\rho \phi$
---	---

**Theorem 5** (Dynamic axiomatization) *The axiomatic system given by Definition 9 and the reduction axioms of Theorem 4 is sound and complete with respect to non-standard ALPSs.*

### 4 Bounded Inference and the Dynamics of interaction

In the previous sections, we have focused on how deductive reasoning, and the bounds imposed on it by cognitive fatigue, affect the agent’s epistemic state. As observed in van Benthem (2008b), apart from “internal elucidation”, external actions such as public announcements (Baltag et al. 1998; Plaza 2007) can also enhance the agent’s epistemic state. The mixed tasks involved in bounded reasoning and in revising epistemic and doxastic states (also discussed in Wassermann 1999) require an account of both sorts of actions and of the ways they are intertwined. The various policies of dynamic change triggered by interaction (public announcement, radical or conservative upgrades, etc., Baltag and Smets 2008; van Benthem 2007) fit in our framework, provided that suitable dynamic operators and model transformations are defined.

#### 4.1 Public Announcements

To supplement the account of a boundedly rational agent who reasons deductively in order to come to know more, we first introduce *public announcements*. These public communication actions can facilitate the agent’s knowledge gain, in this case not because of her own reasoning, but because information was provided to her. For now we assume that the incoming external information was provided to the agent for free (i.e. no cognitive costs are assigned). One common assumption is that such announcements are always truthful and completely trustworthy; the announced sentence is therefore always true and a rational agent always adopts its content.

*Extending the syntax* We introduce operators of the form  $[\psi!]$  to  $\mathcal{L}$ , where  $[\psi!]\phi$  stands for “after announcing  $\psi$ ,  $\phi$  is true”. We focus on cases where  $\psi$  is a propositional formula.<sup>11</sup> Let the language extended with public announcements be called  $\mathcal{L}_{PA}$ .

<sup>11</sup> The case for higher-order announced sentences should be tackled in combination with an account of bounded higher-order reasoning to match the spirit of our framework. More comments on this can be found in Sect. 5.

Extending the semantics  $\mathcal{L}_{PA}$  is interpreted in the plausibility models of Definition 4. The new clause concerns public announcements. Semantically, the formula  $[\psi!]\phi$  is taken to be true in case: whenever  $\psi$  is true,  $\phi$  is true after we eliminate all non- $\psi$  possibilities from  $W$ . This is because the public announcement of  $\psi$  is completely trustworthy, so non- $\psi$  worlds, possible or impossible, are not entertained by the agent any more.<sup>12</sup> The other components of the model, namely  $ord$ ,  $V$ ,  $R$ , are restricted accordingly, while  $cp$  does not change as we view the announcement as provided to the agent externally without any effort on her side. More formally, the update induced by a public announcement is:

**Definition 10** (*Model transformation by public announcement*) Given plausibility model  $M = \langle W^P, W^I, ord, V, R, cp \rangle$ , its transformation by the public announcement of  $\psi \in \mathcal{L}_p$  is the model  $M^{\psi!}$  given by:

$$\left. \begin{aligned} (W^P)^{\psi!} &= \{w \in W \mid M, w \models \psi\} \\ ord^{\psi!} &= ord|_{W^{\psi!}} \\ R^{\psi!} &= R|_{W^{\psi!}} \end{aligned} \right| \begin{aligned} (W^I)^{\psi!} &= \{w \in W \mid M, w \models \psi\} \\ V^{\psi!} &= V|_{W^{\psi!}} \\ cp^{\psi!} &= cp \end{aligned}$$

The conditions of the class **M** are preserved by this definition. The properties of the ordering induced by  $ord$  are guaranteed, just as the public announcement updates preserve the conversely well-founded relation  $\geq$  of the usual plausibility models. *Minimal Consistency* and *Soundness of Rules* still hold, because the worlds surviving the announcement still adhere to these restrictions. *Succession* is preserved because of the way  $W^{\psi!}$  and  $V^{\psi!}$  are defined; if the conditions of *Succession* are met in the updated model, the successor world satisfies  $\psi$  and is therefore included in  $W^{\psi!}$ .

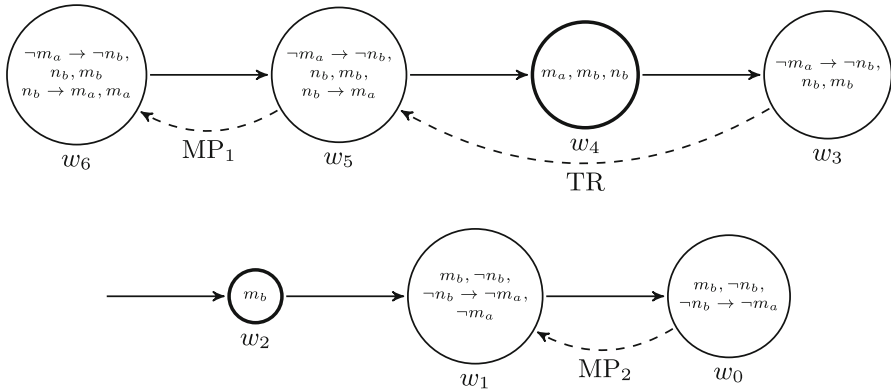
We give the truth clause for public announcements, which follows the standard DEL fashion, only now adapted to the impossible-worlds model we devised to deal with deductive reasoning. In particular, for  $w \in W$ :

$$M, w \models [\psi!]\phi \text{ iff } M, w \models \psi \text{ implies } M^{\psi!}, w \models \phi$$

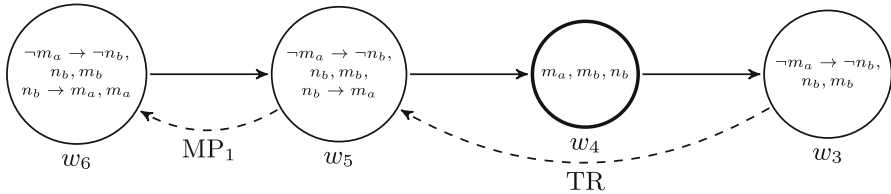
Notice that the formula  $[\psi!]\phi$  is vacuously true if  $\psi$  is not true. The same clause applies to both possible and impossible worlds. This is because of the intuitive interpretation of public announcements. The only worlds surviving the public announcement of  $\psi$  are the ones satisfying  $\psi$ , possible or not, because arguably *any* non- $\psi$  world will be dropped as a possibility.

Under this extended setting, we can bring together external information and the agent's internal reasoning processes. For instance, suppose that the agent knows  $\phi \rightarrow \psi$  and has *MP* available as a rule, and then she comes to know that  $\phi$  from an external source. She may then apply the rule (if affordable) and finally come to know  $\psi$ . It is therefore the combination of interaction and internal deductive reasoning that allowed her to know  $\psi$ . To illustrate the workings of such combinations, we come back to

<sup>12</sup> Notice that in the possible-worlds models of DEL, keeping only the  $\psi$ -satisfying worlds and deleting all the  $\neg\psi$ -satisfying worlds amounts to the same thing due to possible worlds being complete. This is not the case in impossible worlds.



**Fig. 2** The initial model for child *a*, before the announcement of child *b*. As before, thicker borders are used for deductively closed possible worlds. In impossible worlds, we write all propositional formulas satisfied and indicate (non-trivial) expansions via dashed arrows. Notice that the plausibility model continues from  $w_3$  to  $w_2$ .



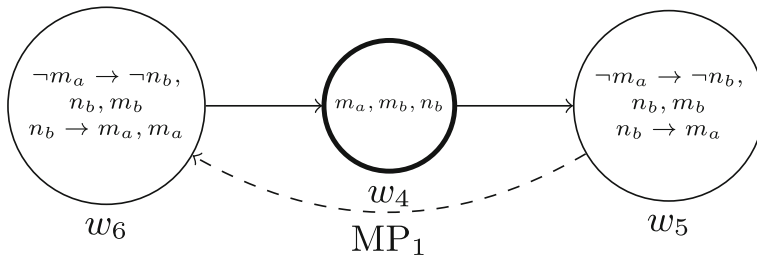
**Fig. 3** The model for child *a* after child *b*'s announcement.

our bounded version of the Muddy Children Puzzle and explicitly account for the interaction taking place.

**Example 2 (Bounded muddy children and public announcements)** In this example, we incorporate the public announcement of child *b* saying no to the question of the father into child *a*'s reasoning process. In particular, *a* before the announcement of *b* cannot tell if she is muddy or not, nor can she figure it out using deductive reasoning alone because her reasoning process depends on the announcement of *b*. We further suppose that the child initially considers it more plausible to be clean. The development of the scenario is presented in Figs. 2, 3, and 4.

*Effortful announcements* We have so far assumed that public announcements are cost-free. However, it can be that adopting a piece of external information requires effort. van Benthem (2008a) and van Benthem (2008b), it is proposed that there are two different kinds of such informational events, presented as “implicit” and “explicit” observations. In our terms, there can be *effortless* announcements (like the ones defined before) and *effortful* announcements.<sup>13</sup> The latter are just like those in Definition 10, but they also incur a cost of accepting the announced information. This presupposes

<sup>13</sup> This also corresponds to a genuine distinction in the philosophical and linguistic literature (Barwise and Perry 1983) between *bare seeing* (“naked infinitives”) and *seeing-that*, which additionally implies epistemic awareness of the fact described.



**Fig. 4** The final model for child *a* after she performs the *TR* inference (as in Example 1), using the information provided by child *b*.

that costs of announcements are also fixed, next to the costs of rules. More specifically, the cost-assigning function  $c$  should be extended as follows:  $c : \mathcal{L}_R \cup \mathcal{L}_p \rightarrow \mathbb{N}^r$ . A simplifying assumption is that announcements of propositional facts always incur the same cost, regardless of the logical structure of the announced sentence. It is nonetheless plausible that the cost of an “explicit” announcement is related to logical complexity, and research on the cognitive difficulty of boolean concepts (Feldman 2000, 2003) might assist in determining these costs. For now we abstain from imposing strict restrictions on  $c$  as the details require empirical evidence and a systematic generalization of it; this is left for further work. In any case, the definition of an effortful announcement is:

**Definition 11** (*Model transformation by an effortful public announcement*) Take plausibility model  $M = \langle W^P, W^I, ord, V, R, cp \rangle$ . Its transformation by the effortful public announcement of  $\psi \in \mathcal{L}_p$  is the model  $M^{\psi!}$  given by:

$$\begin{array}{l|l}
 (W^P)^{\psi!} = \{w \in W \mid M, w \models \psi\} & (W^I)^{\psi!} = \{w \in W \mid M, w \models \psi\} \\
 ord^{\psi!} = ord|_{W^{\psi!}} & V^{\psi!} = V|_{W^{\psi!}} \\
 R^{\psi!} = R|_{W^{\psi!}} & cp^{\psi!} = cp - c(\psi)
 \end{array}$$

The truth clause of an effortful announcement much resembles that of rule-applications. Given that we have terms of the form  $c_\psi$  to express the cost of  $\psi$ :  $M, w \models [\psi!] \phi$  iff  $M, w \models \psi$  implies  $(M, w \models cp \geq c_\psi$  and  $M^{\psi!}, w \models \phi)$

*Reduction* In Sect. 3, we introduced a method to extract a sound and complete axiomatization for our basic framework. This also involved giving reduction axioms for applications of rules. The axiomatization of the Logic of Public Announcement (PAL) (without common knowledge) (Baltag et al. 1998; Baltag and Renne 2016; Plaza 2007) usually involves reduction axioms that allow for replacing formulas with public announcements with—eventually—formulas of the static language. Completeness then follows from the respective complete static base logics. However, the standard reduction axioms (Baltag and Renne 2016; van Benthem 2007; van Ditmarsch et al. 2007) would not work for our purposes. Notice that  $[\psi!] \spadesuit \phi \leftrightarrow (\psi \rightarrow \spadesuit [\psi!] \phi)$ , where  $\spadesuit = K, \Box$ , is valid due to our clause for  $[\psi!] \phi$ . This maintains its intuitive interpretation, also at *impossible* worlds. Despite this validity, a replacement of  $[\psi!] \phi$  in accord with the other reduction axioms would not necessarily go through. Both the



truth clauses for  $K$  and  $\Box$  range over impossible worlds too, to avoid closure under logical equivalence.

In order to reduce formulas with public announcements, we have to follow a procedure similar to the one adopted for rule-applications. That is, we need an auxiliary static operator encoding that  $[\psi!]\phi$  is not evaluated arbitrarily when under the scope of an operator that quantifies over  $W^I$ , instead following the regular public announcement clause.<sup>14</sup>

The addition of a special static operator acting as implication at  $w \in W^I$  is necessary for the reduction of this extended setting. The need for a more expressive language is justified in light of the intuitive readings of  $K$  and  $\Box$ , and their interpretations in Definition 8. Asking that  $K\phi$  and  $\Box\phi$  are true iff  $\phi$  is true throughout the suitable set of possible *and* impossible worlds captures the fallibility of the agent and breaks the forms of logical omniscience. The effect of a truthful public announcement in the agent's epistemic state involves the external information (hence the deletion of worlds), but the prefixed formula is evaluated in the resulting model, which still encodes the limitations of our agent. This is why reducing announcements deviates from the procedure of successive replacements based on the standard reduction axioms. For example,  $[p!]Kp$  (where  $p$  is an atom) is valid because after deleting the non- $p$  worlds,  $Kp$  becomes true. This is equivalent to  $p \rightarrow K[p!]p$  but not to  $p \rightarrow K(p \rightarrow p)$ : the agent does not necessarily know  $p \rightarrow p$ . According to the rationale of our framework, a fallible but bounded agent might have to *reason* to reach  $p \rightarrow p$  too; this piece of knowledge should not be taken for granted.

## 4.2 Other Policies of Integrating External Information

Public announcements are not the only operations for integrating external information. Plausibility models allow us to encode more nuanced notions of knowledge and belief, thus more nuanced policies of integrating external information. For example, the agent might get information coming from a reliable, but not absolutely trustworthy source. This “soft” information, contrary to the “hard” information of a public announcement, triggers a re-arrangement of plausibility, and not an elimination of worlds. Examples of such operations include *radical* (or *lexicographic*) *upgrades* and *conservative upgrades* (Baltag and Renne 2016; Baltag and Smets 2008; van Benthem 2007, 2008b, 2011; Rott 1989). A radical upgrade with  $\psi$  changes the plausibility as follows:  $\psi$ -worlds are ranked over the non- $\psi$  worlds but the ranking of worlds within the two zones remains intact. Regarding conservative upgrades: the most plausible of the  $\psi$ -worlds are ranked over all other worlds and the rest remain unchanged.<sup>15</sup> In

<sup>14</sup> For example, consider a static operator  $[\psi]$  such that  $M, w \models [\psi]\phi$  iff  $M, w \models \psi$  implies  $M, w \models \phi$ , for any  $w \in W$ —this simply amounts to implication if  $w \in W^P$ , but for  $w \in W^I$ , this operator forces us to evaluate the formula classically. Therefore the formula  $[\psi!]I_{\blacklozenge}\phi \leftrightarrow (\psi \rightarrow I_{\blacklozenge}[\psi]\phi)$  is valid and captures the special reading of  $[\psi!]$ , when combined with a  $W^I$ -quantifying operator. It is in terms of this formula, and the one concerning  $L_{\blacklozenge}$ , which behaves normally, that  $[\psi!]_{\blacklozenge}\phi$  can be reduced. An alternative procedure to this is mentioned in Sect. 5.

<sup>15</sup> In fact, a conservative upgrade with  $\psi$  amounts to a radical upgrade with  $best(\psi)$  where  $best(\psi)$  is true at a world iff  $\psi$  is true there, and not true at all worlds strictly more plausible than that (van Benthem 2011).

what follows, we spell out how radical upgrades can be incorporated in our framework, and we notice that more conservative policies can be dealt with along similar lines.

*Extending the syntax* We further expand the language  $\mathcal{L}_{PA}$ , using operators of the form  $[\psi \uparrow]$ , where  $\psi \in \mathcal{L}_p$ , to denote radical upgrades with  $\psi$ . More specifically,  $[\psi \uparrow]\phi$  reads “after the radical upgrade with  $\psi$ ,  $\phi$  is true”.

*Extending the semantics* We hereby present the model transformation by a radical upgrade—again, assuming it amounts to an effortless process.<sup>16</sup> As an auxiliary step take:  $\geq^{\psi \uparrow} = (\geq \cap (W \times [[\psi]])) \cup (\geq \cap ([[ \neg \psi ]] \times W)) \cup (\sim \cap ([[ \neg \psi ]] \times [[\psi]]))$ , where  $[[\psi]]$  denotes the set of worlds where  $\psi$  is not satisfied. Then:

**Definition 12** (*Model transformation by a radical upgrade*) Take plausibility model  $M = \langle W^P, W^I, ord, V, R, cp \rangle$ . Its transformation by a radical upgrade with  $\psi \in \mathcal{L}_p$  is the model  $M^{\psi \uparrow} = \langle W^P, W^I, ord^{\psi \uparrow}, V, R, cp \rangle$ , where  $ord^{\psi \uparrow}$  is any function from the set  $\{f : W \rightarrow \Omega \mid \text{for any } w, u \in W : f(w) \geq f(u) \text{ iff } w \geq^{\psi \uparrow} u\}$ .

Notice that *any* ordinal-assigning function that preserves the ordering of  $\geq^{\psi \uparrow}$  works. This is because we are solely interested in the upgrade having its usual qualitative effect: prioritizing  $\psi$ -worlds over non- $\psi$  ones. The properties of our models are clearly preserved. Then for  $w \in W$ , the truth clause for  $[\psi \uparrow]\phi$  is given by:  $M, w \models [\psi \uparrow]\phi$  iff  $M^{\psi \uparrow}, w \models \phi$ .

Therefore, our plausibility models also facilitate the study of more nuanced attitudes and softer update policies. As an example, we consider an alternative version of the Muddy Children Puzzle found in Baltag and Smets (2009). It treats the incoming information not as “hard” information, but as “soft” information (the sources are considered reliable but not absolutely trustworthy).<sup>17</sup>

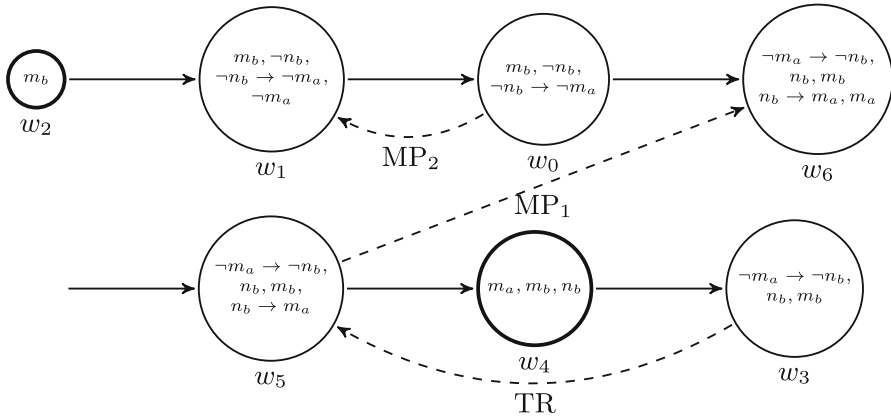
**Example 3** (*Bounded muddy children and radical upgrades*) We now approach the aforementioned scenario of Example 2, taking child  $b$  as a reliable, but not infallible, source of information. Therefore, the incoming information that  $n_b$  is treated as an upgrade, that alters the plausibility ordering, and not as a public announcement, that deletes non- $n_b$  worlds altogether (Fig. 5).

## 5 Conclusions and Further Research

By combining DEL and an impossible-worlds semantics, we modelled fallible but boundedly rational agents who can in principle eliminate their ignorance as long as the task lies within cognitively allowed applications of inference rules. We discussed how this framework accommodates epistemic scenarios realistically and how it fits in the landscape of similar attempts put against logical omniscience. It was shown that this combination can be reduced to a syntactic, possible-worlds structure that allows

<sup>16</sup> Defining *effortful* upgrades, in the spirit explained before regarding public announcements, is also possible.

<sup>17</sup> Other variants of the Puzzle, where two children stand *in line*, therefore one cannot see the other, also show how “softer” notions are better accommodated by plausibility orderings. This, in combination with the focus on boundedly rational agents, demonstrates the salient use of our plausibility models.



**Fig. 5** The model of Fig. 2 after the upgrade with  $n_b$ . Clearly,  $\Box n_b$  is satisfied at the actual world ( $w_4$ ), unlike  $\Box n_b$ . Provided that  $TR$  and  $MP_1$  are affordable and available,  $\langle TR \rangle (MP_1) \Box m_a$  is also satisfied, unlike  $\langle TR \rangle (MP_1) K m_a$ .

for useful formal results. We finally furnished this framework with actions for external information to better account for the fine and mixed nature of reasoning processes.

Note that while factivity of knowledge is indeed warranted by the reflexivity of our models, the correspondence between other properties (such as transitivity) and forms of introspection is disrupted by the impossible worlds. Avoiding unlimited introspection falls within our wider project to model non-ideal agents. Just as with factual reasoning though, we propose a principle of moderation, achieved via the introduction of effortful introspective rules whose semantic effect is similarly projected on the structure of the model. Furthermore, it is precisely along these lines that a multi-agent extension of this setting can be pursued.

Moreover, it is interesting to search for alternatives to the use of special operators in providing reduction axioms for rule-applications and announcements. This might be especially useful for multi-agent frameworks. In particular, there are other tools from DEL that allow uniform treatment of (communicative) actions, such as *action models* (Baltag et al. 1998). Given ongoing work, we believe that action models with postconditions (van Benthem et al. 2006), along with the *set-expressions* used in Velázquez-Quesada (2011, Chapter 5) to embed these into awareness frameworks, could help in obtaining simpler reduction axioms for both rule applications and communicative actions.

Apart from extending the *logical* machinery in order to capture richer reasoning processes, another natural development is towards fine-tuning elements of the model hitherto discussed, in order to better align it with the experimental findings in the literature on rule-based human reasoning. We have already indicated that the function  $c$ , which is responsible for the assignment of cognitive costs, should be sensitive to both the rule-schemas in question and the complexity of their particular instances. The *well-ordering of inferences* that Cherniak (1986) suggests, is supported by the literature we referred to so far, but, at this stage, the evidence fits a qualitative ordering of schemas while a precise quantitative assignment calls for more empirical input. Specifying the

intuitive assumption that the more complex an instance, the more cognitively costly it is, breaks into two tasks (i) choosing some appropriate measure of logical complexity: number of literals, (different) atoms, connectives, etc., (ii) using experimental data to fix coefficients that associate the measure with the performance of agents (with respect to our selected resources). Such a procedure will be pursued in a future paper and it can illuminate whether there are classes of inferences, sharing properties in terms of our measure, that should be assigned equal cognitive costs, as one might intuitively expect.

**Acknowledgements** I would like to thank Sonja Smets for comments and discussions on the contents of the paper. Parts of this paper have been presented at the Logic and Interactive Rationality seminar and the Logic of Conceivability seminar in Amsterdam, as well as at WoLLIC 2018 in Bogotá. I am grateful to the audiences for their very helpful remarks, and especially to Johan van Benthem for comments that have inspired Sect. 4. I am also grateful to the reviewers of WoLLIC 2018 and of this journal for their comments and suggestions.

**Funding** Anthia Solaki was supported by the Netherlands Organization for Scientific Research (project number 322-20-018). The article does not contain any studies with human participants or animals performed by the author.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Appendix

Due to the construction of awareness-like functions, properties of the original model, concerning *Soundness of Rules*, *Minimal Consistency*, and *Succession* are inherited by the reduced model. Clearly, the new quantification sets are  $Q_K(w) = W$  and  $Q_{\square}(w) = \{u \in W \mid w \geq u\}$ . The semantic clauses, based on  $\mathbf{M}$ , are standard for the boolean connectives; the remaining are given below:

- $\mathbf{M}, w \models p$  iff  $p \in V(w)$
- $\mathbf{M}, w \models z_1s_1 + \dots + z_ns_n \geq c$  iff  $z_1s_1^{\mathbf{M}} + \dots + z_ns_n^{\mathbf{M}} \geq c$
- $\mathbf{M}, w \models L_{\clubsuit}\phi$  iff  $\mathbf{M}, u \models \phi$  for all  $u \in Q_{\clubsuit}(w)$
- $\mathbf{M}, w \models I_{\clubsuit}\phi$  iff  $\phi \in I_{\clubsuit}(w)$
- $\mathbf{M}, w \models \spadesuit\phi$  iff  $\mathbf{M}, w \models L_{\clubsuit}\phi$  and  $\mathbf{M}, w \models I_{\clubsuit}\phi$
- $\mathbf{M}, w \models A\rho$  iff  $\rho \in R(w)$
- $\mathbf{M}, w \models \hat{I}_{\clubsuit}\phi$  iff  $\phi \in \hat{I}_{\clubsuit}(w)$
- $\mathbf{M}, w \models \langle RAD \rangle_{\rho}\phi$  iff  $\mathbf{M}, w \models \phi$

Proof for Theorem 1:

**Proof** The proof goes by induction on the complexity of  $\phi$ . Recall that validity is defined with respect to the possible worlds in the original model.

- For  $\phi := p$ :  $\mathbf{M}, w \models p$  iff  $p \in V(w)$  iff  $p \in V(w)$  iff  $\mathbf{M}, w \models p$ .

- For inequalities,  $\neg$ ,  $\wedge$  and  $A$ , the claim is straightforward.
- For  $L_{\clubsuit}\psi: M, w \models L_{\clubsuit}\psi$  iff for all  $u \in W^P \cap Q_{\clubsuit}(w): M, u \models \psi$  iff (by I.H.) for all  $u \in W \cap Q_{\clubsuit}(w): \mathbf{M}, u \models \psi$  iff  $\mathbf{M}, w \models L_{\clubsuit}\psi$ .
- For  $\phi := I_{\clubsuit}\psi: M, w \models I_{\clubsuit}\psi$  iff for all  $u \in W^I \cap Q_{\clubsuit}(w): M, u \models \psi$  iff for all  $u \in W^I \cap Q_{\clubsuit}(w): \psi \in V^+(u)$  iff  $\psi \in I_{\clubsuit}(w)$  iff  $\mathbf{M}, w \models I_{\clubsuit}\psi$ .
- For  $\phi := \spadesuit\psi: M, w \models \spadesuit\psi$  iff for all  $u \in Q_{\clubsuit}(w): M, u \models \psi$ . Since  $u \in W^P \cup W^I$ , this is the case iff  $M, w \models L_{\clubsuit}\psi$  and  $M, w \models I_{\clubsuit}\psi$ . Given the previous steps of the proof, this is the case iff  $\mathbf{M}, w \models L_{\clubsuit}\psi$  and  $\mathbf{M}, w \models I_{\clubsuit}\psi$ , iff  $\mathbf{M}, w \models \spadesuit\psi$ .
- For  $\phi := \hat{I}_{\clubsuit}\psi: M, w \models \hat{I}_{\clubsuit}\psi$  iff for some  $u \in W^I \cap Q_{\clubsuit}(w): M, u \models \psi$  iff for some  $u \in W^I \cap Q_{\clubsuit}(w): \psi \in V^+(u)$  iff  $\psi \in \hat{I}_{\clubsuit}(w)$  iff  $\mathbf{M}, w \models \hat{I}_{\clubsuit}\psi$ .
- For  $\phi := \langle RAD \rangle_{\rho}\psi: M, w \models \langle RAD \rangle_{\rho}\psi$  iff for some  $u \in w^{\rho}: M, u \models \psi$  iff (by I.H. and  $w^{\rho} = \{w\}$  since  $w \in W^P$ )  $\mathbf{M}, w \models \psi$  iff  $\mathbf{M}, w \models \langle RAD \rangle_{\rho}\psi$ .  $\square$

Proof for Theorem 2:

**Proof** Standard arguments suffice regarding *PC*, *Ineq*,  $L_K$ ,  $L_{\square}$  and *Modus Ponens*, *Necessitation<sub>K</sub>*, *Necessitation<sub>\square</sub>* preserve validity as usual. The axioms for *Soundness of Rules*, *Minimal Consistency*, *Succession* are valid due to the respective conditions placed on the model. The validity of *Local Connectedness* is due to the connectedness of the model. The validity of *Indefeasibility*, *Red<sub>\clubsuit</sub>*, *Radius<sub>2</sub>* is a direct consequence of the semantic clauses for  $\spadesuit$ ,  $L_{\clubsuit}$ ,  $I_{\clubsuit}$ . *Radius<sub>1</sub>* is valid due to the deductive closure of possible worlds.<sup>18</sup>  $\square$

Proof for Theorem 3:

Towards showing completeness, we use a suitable canonical model. Taking (maximal)  $\Lambda$ -consistent sets and showing Lindenbaum’s lemma follow the standard paradigm.

**Definition 13** (*Canonical model*) The canonical model for the logic  $\Lambda$  is  $\mathcal{M} := \langle \mathcal{W}, \geq, \sim, \mathcal{V}, \mathcal{R}, cp, \mathcal{I}_{\clubsuit}, \hat{\mathcal{I}}_{\clubsuit} \rangle$  where:

- $\mathcal{W}$  is the set of all maximal  $\Lambda$ -consistent sets.
- $\geq$  is such that for  $w, u \in \mathcal{W}: w \geq u$  iff  $\{\phi \mid L_{\square}\phi \in w\} \subseteq u$ .
- $\sim$  is such that for  $w, u \in \mathcal{W}: w \sim u$  iff  $\{\phi \mid L_K\phi \in w\} \subseteq u$ .
- $\mathcal{V}(w) = \{p \mid p \in w\}$ , with  $w \in \mathcal{W}$ .
- $\mathcal{R}(w) = \{\rho \mid A\rho \in w\}$ , with  $w \in \mathcal{W}$ .
- $\mathcal{I}_{\clubsuit}(w) = \{\phi \mid I_{\clubsuit}\phi \in w\}$ , with  $w \in \mathcal{W}$ .
- $\hat{\mathcal{I}}_{\clubsuit}(w) = \{\phi \mid \hat{I}_{\clubsuit}\phi \in w\}$ , with  $w \in \mathcal{W}$ .

Due to  $L_K$ ,  $L_{\square}$ , *Indefeasibility*, *Local Connectedness* and modal logic results on correspondence (Blackburn et al. 2001) the canonical model is reflexive, transitive and (locally) connected (with respect to  $\geq$ ) and  $\sim$  is the symmetric extension of  $\geq$  (these properties yield the so-called *non-standard* plausibility models). The axioms on *Soundness of Rules*, *Minimal Consistency*, *Succession* and *Radius* are such to ensure that the model has the desired properties.

<sup>18</sup> Notice that the fact that the interpretations of  $\langle RAD \rangle_{\rho}$  and  $[RAD]_{\rho}$  are not arbitrary in impossible worlds is important in this proof.

We then perform induction on the complexity of  $\phi$  to show the truth lemma:  $\mathcal{M}, w \models \phi$  iff  $\phi \in w$ . The claim for propositional atoms, the boolean cases, linear inequalities, and  $A$  holds, due to the construction of the canonical model (namely,  $\mathcal{V}$  and  $\mathcal{R}$ ), *Ineq*, the properties of maximal consistent sets and I.H.. The claim for  $\langle \text{RAD} \rangle_\rho$  follows by the I.H. and *Radius*<sub>1</sub>. The claims for  $L_\square$  and  $L_K$  follow with the help of the I.H. as usual in Modal Logic, because these operators behave as normal normal operators (Blackburn et al. 2001). For  $I_\square$ ,  $I_K$ ,  $\hat{I}_\square$  and  $\hat{I}_K$  we rely on the construction of the awareness-like functions and then the result is immediate. For  $K\phi$  and  $\square\phi$ , we make use of *Red*<sub>K</sub> and *Red* <sub>$\square$</sub> , the I.H. and the results of the previous steps on  $L_K$ ,  $I_K$  and  $L_\square$ ,  $I_\square$ .<sup>19</sup>

Proof for Theorem 4:

**Proof** – The claim is easy for the atoms, the boolean cases, the inequalities,  $A$ ,  $\langle \text{RAD} \rangle_\rho$ ,  $L_K$  and  $L_\square$ . We will only show why the claim holds for  $I_K$  and  $I_\square$ ,  $\hat{I}_K$  and  $\hat{I}_\square$  because the claims involving  $K$ ,  $\square$  will then follow from the clause for  $\spadesuit$  and the distribution of  $\langle \rho \rangle$  over conjunction.

- Let  $M$  be an arbitrary model and  $w$  an arbitrary possible world of the model. Suppose  $M, w \models \langle \rho \rangle I_K \phi$ . Therefore  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$  and  $M^\rho, w \models I_K \phi$ . Recall that  $W^\rho = \bigcup_{u \in W} u^\rho$ . Therefore for all  $v \in W^\rho \cap W^I$ ,  $M^\rho, v \models \phi$  (1). Take arbitrary  $u \in W^I$  and arbitrary  $v \in u^\rho$ . Then,  $v \in W^\rho \cap W^I$ , and by (1) and the definitions of  $V$  and radius:  $M, v \models \phi$ . Overall,  $M, w \models I_K[\text{RAD}]_\rho \phi$  and by  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$ , we finally get  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_K[\text{RAD}]_\rho \phi$ . For the other direction, suppose that  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_K[\text{RAD}]_\rho \phi$ . Take arbitrary  $v \in W^\rho \cap W^I$ , i.e. there is some  $u \in W^I$  such that  $v \in u^\rho$ . By the truth conditions of  $M, w \models I_K[\text{RAD}]_\rho \phi$ , for all  $u \in W^I$ ,  $M, u \models [\text{RAD}]_\rho \phi$ , i.e. for all  $v \in u^\rho$ :  $M, v \models \phi$ . Therefore, for our arbitrary  $v$ , it is the case that  $M, v \models \phi$ , and by definitions of  $V$  and radius,  $M^\rho, v \models \phi$ . Overall,  $M^\rho, w \models I_K \phi$  and finally  $M, w \models \langle \rho \rangle I_K \phi$ .
- Let  $M$  be an arbitrary model and  $w$  an arbitrary possible world of the model. Suppose  $M, w \models \langle \rho \rangle I_\square \phi$ . Therefore  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$  and  $M^\rho, w \models I_\square \phi$ . Since  $W^\rho = \bigcup_{u \in W} u^\rho$ , for all  $v \in W^\rho \cap W^I$  such that  $w \geq^\rho v$ :  $M^\rho, v \models \phi$  (1). Then, take arbitrary  $u \in W^I \cap Q_\square(w)$  and arbitrary  $v \in u^\rho$ . Since  $\text{ord}^\rho(v) \leq \text{ord}(u)$  (by Step 2 of transformation) and  $w \geq u$ , we get that  $w \geq^\rho v$ . Therefore  $v \in W^\rho \cap W^I$ , and by (1) and the definitions of  $V$  and radius:  $M, v \models \phi$ . Hence  $M, w \models I_\square[\text{RAD}]_\rho \phi$  and by  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$ , we finally get  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_\square[\text{RAD}]_\rho \phi$ . For the other direction, suppose that  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_\square[\text{RAD}]_\rho \phi$ . Take arbitrary  $v \in W^\rho \cap W^I$  such that  $w \geq^\rho v$ , i.e. there is some  $u \in W^I$  such that  $v \in u^\rho$ . Take the most plausible of these worlds (from which  $v$  originated). For this  $u$ , since  $\text{ord}^\rho(v) = \text{ord}(u)$  and  $w \geq^\rho v$  then  $w \geq u$ . By the truth conditions of  $M, w \models I_\square[\text{RAD}]_\rho \phi$ , for all  $u \in W^I$  such that  $w \geq u$ :  $M, u \models [\text{RAD}]_\rho \phi$ , i.e. for all  $v \in u^\rho$ :  $M, v \models \phi$ .

<sup>19</sup> In fact, we can claim that this logic is weakly complete with respect to ALPSs where  $\geq$  is conversely well-founded. This is because our structures have the finite model property (as the established plausibility models, e.g. in Baltag and Smets (2008); Velázquez-Quesada (2011)), so there are no infinite  $>$  chains of more and more plausible worlds.

- Therefore, for our arbitrary  $v$ , it is the case that  $M, v \models \phi$ , and by definitions of  $V$  and radius,  $M^\rho, v \models \phi$  too. Overall,  $M^\rho, w \models I_\square\phi$  and finally  $M, w \models \langle \rho \rangle I_\square\phi$ .
- Let  $M$  be an arbitrary model and  $w$  an arbitrary possible world of the model. Suppose  $M, w \models \langle \rho \rangle \hat{I}_K\phi$ . Therefore  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$  and  $M^\rho, w \models \hat{I}_K\phi$ . Since  $W^\rho = \bigcup_{u \in W} u^\rho$ , for some  $v \in W^\rho \cap W^I$ :  $M^\rho, v \models \phi$ . That is, there is some  $u \in W^I$  such that  $v \in u^\rho$  and  $M^\rho, v \models \phi$ . Therefore, for some  $u \in W^I$  there is some  $v \in u^\rho$  such that (by definitions of  $V$  and radius)  $M, v \models \phi$ . This amounts to  $M, w \models \hat{I}_K \langle RAD \rangle_\rho \phi$ , and overall  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_K \langle RAD \rangle_\rho \phi$ . For the other direction, suppose  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_K \langle RAD \rangle_\rho \phi$ . From  $M, w \models \hat{I}_K \langle RAD \rangle_\rho \phi$ , we get that there is some  $u \in W^I$  such that for some  $v \in u^\rho$ :  $M, v \models \phi$ . But then  $v \in W^\rho \cap W^I$  and by definitions of  $V$  and radius,  $M^\rho, w \models \hat{I}_K\phi$ . So overall  $M, w \models \langle \rho \rangle \hat{I}_K\phi$ .
  - Let  $M$  be an arbitrary model and  $w$  an arbitrary possible world of the model. Suppose  $M, w \models \langle \rho \rangle \hat{I}_\square\phi$ . Therefore  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho$  and  $M^\rho, w \models \hat{I}_\square\phi$ . Since  $W^\rho = \bigcup_{u \in W} u^\rho$ , for some  $v \in W^\rho \cap W^I$  with  $w \geq^\rho v$ :  $M^\rho, v \models \phi$ . That is, there is  $u \in W^I$  such that  $v \in u^\rho$ , and  $w \geq^\rho v$  and  $M^\rho, v \models \phi$ . Take the most plausible such  $u$ . Since  $ord(u) = ord^\rho(v)$ ,  $w \geq u$ . By these, and definitions of  $V$  and radius: there is  $u \in W^I \cap Q_\square(w)$  and  $v \in u^\rho$  with  $M, v \models \phi$ , which is precisely  $M, w \models \hat{I}_\square \langle RAD \rangle_\rho \phi$ . Overall:  $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_\square \langle RAD \rangle_\rho \phi$ . For the other direction, suppose  $M, w \models (cp \geq c_\rho)$ ,  $M, w \models A\rho \wedge \hat{I}_\square \langle RAD \rangle_\rho \phi$ . This means that there is some  $u \in W^I \cap Q_\square(w)$  and some  $v \in u^\rho$  with  $M, v \models \phi$ . It suffices to show that  $M^\rho, w \models \hat{I}_\square\phi$ . But from  $w \geq u$  and  $v \in u^\rho$ , we obtain that  $w \geq^\rho v$ , and  $v \in W^\rho$ . Due to this and definitions of  $V$  and radius:  $M^\rho, v \models \phi$  and then  $M^\rho, w \models \hat{I}_\square\phi$ . Overall indeed  $M, w \models \langle \rho \rangle \hat{I}_\square\phi$ . □

Proof for Theorem 5:

**Proof** The proof follows from Theorems 2, 3, and 4. □

## References

- Alechina, N., & Logan, B. (2009). A logic of situated resource-bounded agents. *Journal of Logic, Language and Information*, 18, 79–95.
- Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge, TARK'98* (pp. 43–56). Morgan Kaufmann Publishers Inc.
- Baltag, A., & Renne, B. (2016). Dynamic epistemic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edn.
- Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. *Logic and the Foundations of Game and Decision Theory, Texts in Logic and Games*, 3, 9–58.
- Baltag, A., & Smets, S. (2009). Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In *TARK*.
- Bara, B. G., Bucciarelli, M., & Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. *The American Journal of Psychology*, 108(2), 157–193.
- Barwise, J., & Perry, J. (1983). *Situations and attitudes*. MIT Press.
- 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*. Cambridge University Press.

- Cherniak, C. (1986). *Minimal rationality*. Bradford book, MIT Press.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 24, 87–114.
- Ditmarsch, H., Halpern, J. Y., van der Hoek, W., & Kooi, B. P. (2015). *Handbook of epistemic logic*. College Publications.
- Duc, H. N. (1997). Reasoning about rational, but not logically omniscient, agents. *Journal of Logic and Computation*, 7(5), 633. <https://doi.org/10.1093/logcom/7.5.633>.
- Fagin, R., & Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1), 39–76.
- Fagin, R., & Halpern, J. Y. (1994). Reasoning about knowledge and probability. *Journal of the ACM*, 41(2), 340–367.
- Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. MIT Press.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407, 630 EP
- Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, 47(1), 75–89.
- Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4(4), 475–484.
- Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, 99(3), 418–439.
- Kahneman, D., & Beatty, J. (1967). Pupillary responses in a pitch-discrimination task. *Perception & Psychophysics*, 2(3), 101–105.
- Lehrer, K. (2000). *Theory of knowledge*. Westview Press.
- Miller, G. (1956). The magical number seven, plus or minus 2: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Parikh, R. (1987). Knowledge and the problem of logical omniscience. In *Proceedings of the second international symposium on methodologies for intelligent systems* (pp. 432–439). North-Holland Publishing Co.
- Plaza, J. (2007). Logics of public communications. *Synthese*, 158(2), 165–179.
- Sears, C. R., & Pylyshyn, Z. (2000). Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*, 54, 1–14.
- Rantala, V. (1982). Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica*, 35, 106–115.
- Rasmussen, M. S. (2015). Dynamic epistemic logic and logical omniscience. *Logic and Logical Philosophy*, 24, 377–399.
- Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.
- Rott, H. (1989). Conditionals and theory change: Revisions, expansions, and additions. *Synthese*, 81(1), 91–113.
- Schroyens, W., & Schaeken, W. (2003). A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 140–149
- Schroyens, W. J., Schaeken, W., & D'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning*, 7(2), 121–172
- 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). Springer.
- Spohn, W. (1988). Ordinal conditional functions. A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics* (Vol. II). Kluwer Academic Publishers.
- Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, 128(1), 169–199.
- 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*. MIT Press.
- van Benthem, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2), 129–155.
- van Benthem, J. (2008). Merging observation and access in dynamic epistemic logic. *Studies in Logic*, 1, 1–16.
- van Benthem, J. (2008). Tell it like it is: Information flow in logic. *Journal of Peking University (Humanities and Social Science Edition)*, 1, 80–90.



- van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.
- van Benthem, J., van Eijck, J., & Kooi, B. (2006). Logics of communication and change. *Information and Computation*, 204(11), 1620–1662.
- van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic* (1st ed.). Springer Publishing Company, Incorporated.
- Velázquez-Quesada, F. R. (2011). Small steps in dynamics of information. Ph.D. thesis, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands
- Wansing, H. (1990). A general possible worlds framework for reasoning about knowledge and belief. *Studia Logica*, 49(4), 523–539.
- Wassermann, R. (1999). Resource bounded belief revision. *Erkenntnis*, 50(2), 429–446.
- Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences*, 13(4), 167–174.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.