The effort of reasoning: modelling the inference steps of boundedly rational agents

Sonja Smets[1] and Anthia Solaki[1]

ILLC, University of Amsterdam, the Netherlands
{s.j.l.smets,a.solaki2}@uva.nl

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 finally 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.

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 ([14]). Models for 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 reveal that subjects are *systematically* fallible in reasoning tasks ([32, 33]). 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 some 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* ([9]). 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 aim to design a normative model, rather than a purely descriptive one.

As an illustration we consider the standard *Muddy Children Puzzle* ([14]) which is based on the unrealistic assumption that children are unbounded and perfect logicians, who can perform demanding deductive steps 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 [21] 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) ([3, 4, 7, 11]). DEL is equipped with dynamic operators, which can well be used to denote applications of inference rules. We give a semantics of these operators via plausibility models ([4]). 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 ([16]) 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.

An arbitrary syntactic awareness-filter, used to discern *explicit* attitudes as in [13], cannot work for our purposes for it cannot be associated with logical competence, and 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 [34]; 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 [1, 25, 26]. 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 combines the benefits of plausibility models and impossible worlds in the realistic modelling of competent but bounded reasoners, and also suggests how their technical treatment can be facilitated.

The paper is structured as follows: in Section 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 Section 3, followed by our conclusions and directions for further work in Section 4.

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 Φ. Using this notation we first define *inference rules*:

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

We use $pr(\rho)$ and $con(\rho)$ to abbreviate, respectively, the set of premises and the conclusion of a rule ρ 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 ρ is given by the following implication in $\mathcal{L}_p$, i.e. $tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \to con(\rho)$*

We introduce the language $\mathcal{L}$, extending $\mathcal{L}_p$ with two epistemic modalities: K for conventional knowledge, and $\Box$ for *defeasible knowledge.* As argued in [4], it is philosophically interesting to include both attitudes in one system. While K represents an agent's full introspective and factive attitude, $\Box$ is factive but not fully introspective. This weaker notion satisfies the S4-properties and is inspired by the defeasibility analysis of knowledge ([19, 31]), 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 labeled 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_ρ of inference rules $\rho \in \mathcal{L}_R$, and the cognitive capacity cp. Given a set of propositional atoms Φ, the language $\mathcal{L}$ is defined by:*

$$\phi ::= p \mid z_1 s_1 + \ldots + z_n s_n \geq \mathrm{c} \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid \Box\phi \mid A\rho \mid \langle\rho\rangle\phi$$

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

The language comprises linear inequalities of the form $z_1 s_1 + \ldots + z_n s_n \geq \mathrm{c}$, to deal with cognitive effort via comparisons of costs and capacity.[1] The modalities K and $\Box$ represent infallible and defeasible knowledge, respectively. The operator A indicates the agent's availability of inference rules, i.e. $A\rho$ denotes that the agent has acknowledged rule ρ 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 ρ, ϕ 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 $s_1 - s_2 \geq \overline{0}$.

2.2 Plausibility models

Our semantics is based on plausibility models ([4]). In line with [30] we use a mapping from the set of worlds to the class of ordinals Ω to derive the plausibility ordering. The model is augmented by impossible worlds, which 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 attempt. 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: [17, 27, 33] 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.[2]

[1] Notice that c is an r-tuple. The choice of r is discussed in the next subsection.

[2] 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

Every model that we consider 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.

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* ([10, 20]).[3]

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* Ω *assigning an ordinal to each world.* $V : W \to \mathcal{P}(\mathcal{L})$ *is a valuation function mapping each world to a set of formulas.* $R : W \to \mathcal{P}(\mathcal{L}_R)$ *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.*

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.[4] 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 being "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.[5] We define $\sim$, representing epistemic indistinguishability: $w \sim u$ iff either $w \geq u$ or $u \geq w$.

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* $w \in W^I$*,* $V^*(w) = \{\phi \in \mathcal{L}_p \mid \phi \in V(w)\}$*.*

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.

[3] Numerical assignments might also pertain to the use of pupil assessment and eye-tracking as measures of attention and indicators of cognitive effort [18, 23, 37].

[4] 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.

[5] 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.

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 (ρ-radius). *The ρ-radius of a world w is given by* [6]*:*

$$w^{\rho} := \begin{cases} \{w\}, \ \textit{if } pr(\rho) \nsubseteq V^*(w) \\ \emptyset, \ \textit{if } pr(\rho) \subseteq V^*(w) \ \textit{and } (\neg con(\rho) \in V^*(w) \ \textit{or } con(\rho) = \neg\phi \\ \textit{for some } \phi \in V^*(w)) \\ \{u \mid u \textit{ the successor of } w\}, \textit{if } pr(\rho) \subseteq V^*(w) \ \textit{and } \neg con(\rho) \notin V^*(w) \\ \ \textit{and } con(\rho) \neq \neg\phi \textit{ for all } \phi \in V^*(w) \end{cases}$$

The ρ-radius, inspired by [26], represents how the rule ρ triggers an informational change and its element, if it exists, is called *ρ-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 ρ-expansion is unique. As ρ-expansions 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 ρ-expansion amounts to itself for $w \in W^P$ (due to the deductive closure of possible worlds) while an impossible world's ρ-expansion is another impossible world.

2.3 Model transformations and semantic clauses

To evaluate $\langle\rho\rangle\phi$, we have to examine the truth value of ϕ in a transformed model, defined in such a way to capture the effect of applying ρ. Roughly, a pointed plausibility model (M', w) (which consists of a plausibility model and a point indicating the real world) is the *ρ-update* of a given pointed plausibility model, whenever the set of worlds is replaced by the worlds reachable by an application of ρ 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 ρ 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:

Step 1 Given a rule ρ, $W^{\rho} := \bigcup_{v \in W} v^{\rho}$. In words, the ρ-expansions of the worlds initially entertained by the agent. So the ρ-updated pointed model (M^{ρ}, w) should be such that its set of worlds is W^{ρ}. As observed above, any elimination of worlds is in fact an elimination affecting the set W^I.

[6] 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.

Step 2 We now develop the new ordering ord^{ρ} following the application of the inference rule. Let $u \in W^{\rho}$. 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^{ρ}, 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^{ρ} 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_{ρ} 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 ϕ is true at w in M (notation: $M, w \models \phi$) and when ϕ is false at w in M (notation: $M, w \Dashv \phi$). For $w \in W^{I}$: $M, w \models \phi$ iff $\phi \in V(w)$, and $M, w \Dashv \phi$ iff $\neg\phi \in V(w)$. For $w \in W^{P}$, given that the boolean cases are standard:*

<table>
<tr><td>$M, w \models p$ <i>iff</i> $p \in V(w)$, <i>where</i> $p \in \Phi$</td><td>$M, w \Dashv \phi$ <i>iff</i> $M, w \not\models \phi$</td></tr>
<tr><td>$M, w \models K\phi$ <i>iff</i> $M, u \models \phi$ <i>for all</i> $u \in W$</td><td>$M, w \models A\rho$ <i>iff</i> $\rho \in R(w)$</td></tr>
<tr><td colspan="2">$M, w \models \Box\phi$ <i>iff</i> $M, u \models \phi$ <i>for all</i> $u \in W$ <i>such that</i> $w \geq u$</td></tr>
<tr><td colspan="2">$M, w \models \langle\rho\rangle\phi$ <i>iff</i> $M, w \models (cp \geq c_{\rho})$, $M, w \models A\rho$ <i>and</i> $M^{\rho}, w \models \phi$</td></tr>
<tr><td colspan="2">$M, w \models z_1 s_1 + \ldots + z_n s_n \geq \mathrm{c}$ <i>iff</i> $z_1 s_1^{M} + \ldots + z_n s_n^{M} \geq \mathrm{c}$</td></tr>
</table>

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 ρ are sensitive to the idea of resource-boundedness, unlike plain assertions. The latter require that ϕ 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 ϕ follows from the accessible worlds via ρ. 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 ϕ 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 [12, 26] 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.[7] Furthermore, the enterprise of providing a semantics contributes to [25]'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 attempt 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 $TR = \{\neg m_a \to \neg n_b\} \rightsquigarrow n_b \to m_a$, $MP = \{n_b, n_b \to m_a\} \rightsquigarrow m_a$, $Res = \{time, memory\}$, $c(TR) = (5, 2), c(MP) = (2, 2)$, $cp = (5, 7)$.

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 $\Box(\neg m_a \to \neg n_b)$ 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

[7] See, for example, [9, 17, 27, 28, 29]. 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.

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 $\langle TR \rangle \Box (n_b \to m_a)$ 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.

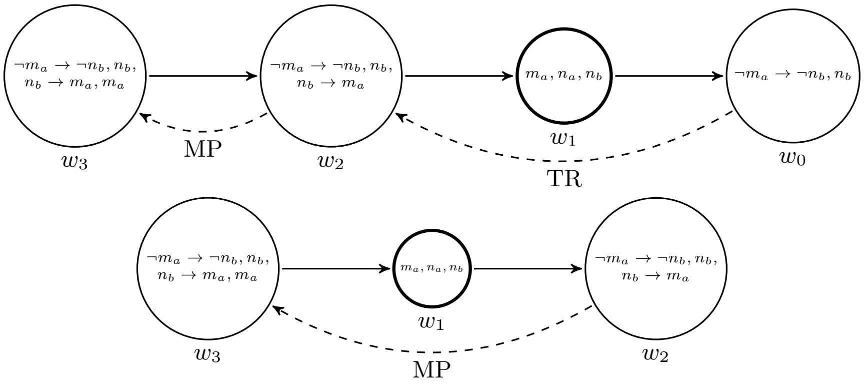
$\neg m_a \to \neg n_b, n_b,$
$n_b \to m_a, m_a$
w_3
MP
$\neg m_a \to \neg n_b, n_b,$
$n_b \to m_a$
w_2
m_a, n_a, n_b
w_1
TR
$\neg m_a \to \neg n_b, n_b$
w_0
$\neg m_a \to \neg n_b, n_b,$
$n_b \to m_a, m_a$
w_3
m_a, n_a, n_b
w_1
MP
$\neg m_a \to \neg n_b, n_b,$
$n_b \to m_a$
w_2

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.

3 Reduction and axiomatization

Work in [35] shows how various models for knowledge and belief, including structures for awareness ([13]), can be viewed as impossible-worlds models (more specifically, *Rantala models*, [24]), 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; 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 construct a *canonical model* to obtain a sound and complete axiomatization for the static part. 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 $\Box$ and take the *quantification set* $Q_\spadesuit(w)$ to be W if $\spadesuit = K$, and $Q_\spadesuit(w) := \{u \mid w \geq u\}$, if $\spadesuit = \Box$ (to denote the set that the truth clauses for K and $\Box$ 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 words, for each $\spadesuit = K, \Box$. 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 $\langle RAD \rangle_\rho$, labeled by inference rules are such to ensure that there is some ϕ-satisfying ρ-expansion. To express that *all* ρ-expansions are ϕ-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 ϕ is of the form $\neg\psi$, for some formula ψ, then $\overline{I}_\spadesuit \phi := \hat{I}_\spadesuit \psi$, else $\overline{I}_\spadesuit \phi := \bot$.

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:

- $\mathrm{I}_\spadesuit : W^P \rightarrow \mathcal{P}(\mathcal{L}_r)$ such that $\mathrm{I}_\spadesuit(w) = \bigcap\limits_{v \in W^I \cap Q_\spadesuit(w)} V^+(v)$. Intuitively, $\mathrm{I}_\spadesuit$ takes a possible world w and yields the set of those formulas that are true at all impossible worlds in its quantification set.
- $\hat{\mathrm{I}}_\spadesuit : W^P \rightarrow \mathcal{P}(\mathcal{L}_r)$ such that $\hat{\mathrm{I}}_\spadesuit(w) = \bigcup\limits_{v \in W^I \cap Q_\spadesuit(w)} V^+(v)$. Intuitively, $\hat{\mathrm{I}}_\spadesuit$ 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 $\mathbf{M} = \langle \mathrm{W}, \geq, \sim, \mathrm{V}, \mathrm{R}, cp, \mathrm{I}_{\spadesuit}, \hat{\mathrm{I}}_{\spadesuit}\rangle$ where:

$\mathrm{W} = W^P$ $\qquad$ $\mathrm{V}(w) = V(w)$ for $w \in \mathrm{W}$

$u \geq w$ iff $ord(u) \geq ord(w)$, for $w, u \in \mathrm{W}$ $\qquad$ $\mathrm{R}(w) = R(w)$ for $w \in \mathrm{W}$

$u \sim w$ iff $u \geq w$ or $w \geq u$, for $w, u \in \mathrm{W}$ $\qquad$ $\mathrm{I}_{\spadesuit}, \hat{\mathrm{I}}_{\spadesuit}$ as explained before

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 $\mathbf{M}, w \models \phi$.*

3.3 Axiomatization

We have reduced plausibility models to ALPS. We now develop the (static) logic Λ, showing that it is sound and complete w.r.t. them.

Definition 9 (Axiomatization of Λ). *Λ is axiomatized by Table 1 and the rules Modus Ponens, Necessitation$_K$ (from ϕ, infer $L_K\phi$) and Necessitation$_\Box$ (from ϕ, infer $L_\Box\phi$).*

Table 1.

Axioms			
PC	All instances of classical propositional tautologies		
Ineq	All instances of valid formulas about linear inequalities		
L_K	The S5 axioms for L_K	*Soundness of Rules*	$A\rho \to tr(\rho)$
$L_\Box$	The S4 axioms for $L_\Box$	*Minimal Consistency*	$\neg(I_\Box\phi \wedge I_\Box\neg\phi)$
Succession$_1$	$(\bigwedge_{\psi\in pr(\rho)} I_{\spadesuit}\psi \wedge \neg\hat{I}_{\spadesuit}\neg con(\rho) \wedge \neg\overline{I}_{\spadesuit} con(\rho)) \to$ $I_{\spadesuit}\langle RAD\rangle_\rho con(\rho) \wedge (I_{\spadesuit}\phi \to I_{\spadesuit}\langle RAD\rangle_\rho\phi)$, for $\phi \in \mathcal{L}_p$		
Succession$_2$	$I_{\spadesuit}\langle RAD\rangle_\rho\phi \to I_{\spadesuit}\phi$, for $\phi \in \mathcal{L}_p$ and $\phi \neq con(\rho)$		
Succession$_3$	$\neg\bigwedge_{\psi\in pr(\rho)} I_{\spadesuit}\psi \to (I_{\spadesuit}\phi \leftrightarrow I_{\spadesuit}\langle RAD\rangle_\rho\phi)$, for $\phi \in \mathcal{L}_p$		
Succession$_4$	$\bigwedge_{\psi\in pr(\rho)} I_{\spadesuit}\psi \wedge (\hat{I}_{\spadesuit}\neg con(\rho) \vee \overline{I}_{\spadesuit} con(\rho)) \to I_{\spadesuit}[RAD]_\rho\bot$		
Local Connectedness	$L_K(\phi \vee L_\Box\psi) \wedge L_K(\psi \vee L_\Box\phi) \to (L_K\phi \vee L_K\psi)$		
Red$_{\spadesuit}$	$\spadesuit\phi \leftrightarrow (L_{\spadesuit}\phi \wedge I_{\spadesuit}\phi)$	*Indefeasibility*	$L_K\phi \to L_\Box\phi$
Radius$_1$	$\langle RAD\rangle_\rho\phi \leftrightarrow \phi$		$I_K\phi \to I_\Box\phi$
Radius$_2$	$I_{\spadesuit}[RAD]_\rho\phi \leftrightarrow (I_{\spadesuit}\langle RAD\rangle_\rho\top \to I_{\spadesuit}\langle RAD\rangle_\rho\phi)$		

Ineq, described in [15], is introduced to accommodate the linear inequalities.[8] The S5 axioms for L_K and S4 axioms for $L_\Box$ mimic the behaviour of K and $\Box$ 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, $Red_\spadesuit$ expresses K and $\Box$ 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). *Λ is sound w.r.t. ALPS.*

Theorem 3 (Completeness). *Λ is complete w.r.t. non-standard* [9] *ALPS.*

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 $c_\rho^\rho := c_\rho$.

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

$$\begin{array}{l|l}
\langle\rho\rangle p \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge p & \langle\rho\rangle\neg\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \neg\langle\rho\rangle\phi \\
\langle\rho\rangle(\phi \wedge \psi) \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \langle\rho\rangle\phi \wedge \langle\rho\rangle\psi & \langle\rho\rangle L_\spadesuit\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge L_\spadesuit\phi \\
\langle\rho\rangle I_\spadesuit\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge I_\spadesuit[RAD]_\rho\phi & \langle\rho\rangle\spadesuit\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \spadesuit[RAD]_\rho\phi \\
\langle\rho\rangle A\sigma \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge A\sigma & \langle\rho\rangle\hat{I}_\spadesuit\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_\spadesuit\langle RAD\rangle_\rho\phi \\
\langle\rho\rangle\langle RAD\rangle_\rho\phi \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge \langle RAD\rangle_\rho\phi &
\end{array}$$

$$\langle\rho\rangle(z_1 s_1 + \ldots + z_n s_n \geq \mathrm{c}) \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge (z_1 s_1^\rho + \ldots + z_n s_n^\rho \geq \mathrm{c})$$

Theorem 5 (Dynamic axiomatization). *The axiomatic system given by Table 1 and the reduction axioms of Theorem 4 is sound and complete w.r.t. non-standard ALPS.*

4 Conclusions and further research

By combining DEL and an impossible-wolds semantics, we modeled 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 finally shown that this combination can be reduced to a syntactic, possible-worlds structure that allows for useful formal results.

[8] *Ineq* is of course slightly adapted as terms are interpreted as r-tuples. This makes no difference for the axioms in [15], with the exception of *dichotomy* which is not needed given our reading of inequality.

[9] More on this terminology can be found in the Appendix.

We have focused on how deductive reasoning affects the agent's epistemic state. As observed in [6], apart from "internal elucidation", external actions, e.g. public announcements ([2, 22]), also enhance the agent's state. The mixed tasks involved in bounded reasoning and in revising epistemic states (also discussed in [36]) 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 upgrades etc., [4, 5]) fit in our framework, provided that suitable dynamic operators and model transformations are defined. For instance, public announcements eliminate the worlds that do not satisfy the announced sentence and radical upgrades prioritize those that satisfy it.

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.

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 [9] 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 (w.r.t. 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.

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 $\mathrm{Q}_K(w) = \mathrm{W}$ and $\mathrm{Q}_\Box(w) = \{u \in \mathrm{W} \mid w \geq u\}$. The semantic clauses, based on $\mathbf{M}$, are standard for the boolean connectives; the remaining are given below:

- $\mathbf{M}, w \models p$ iff $p \in \mathrm{V}(w)$

- $\mathbf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq \mathrm{c}$ iff $z_1 s_1^{\mathbf{M}} + \ldots + z_n s_n^{\mathbf{M}} \geq \mathrm{c}$
- $\mathbf{M}, w \models L_{\spadesuit}\phi$ iff $\mathbf{M}, u \models \phi$ for all $u \in \mathrm{Q}_{\spadesuit}(w)$
- $\mathbf{M}, w \models I_{\spadesuit}\phi$ iff $\phi \in \mathrm{I}_{\spadesuit}(w)$
- $\mathbf{M}, w \models \spadesuit\phi$ iff $\mathbf{M}, w \models L_{\spadesuit}\phi$ and $\mathbf{M}, w \models I_{\spadesuit}\phi$
- $\mathbf{M}, w \models A\rho$ iff $\rho \in \mathrm{R}(w)$
- $\mathbf{M}, w \models \hat{I}_{\spadesuit}\phi$ iff $\phi \in \hat{\mathrm{I}}_{\spadesuit}(w)$
- $\mathbf{M}, w \models \langle RAD \rangle_{\rho}\phi$ iff $\mathbf{M}, w \models \phi$

Proof for Theorem 1:

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

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

Proof for Theorem 2:

Proof. Standard arguments suffice regarding *PC*, *Ineq*, L_K, $L_{\Box}$ and *Modus Ponens*, *Necessitation*$_K$, *Necessitation*$_{\Box}$ 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*$_{\spadesuit}$, *Radius*$_2$ is a direct consequence of the semantic clauses for $\spadesuit, L_{\spadesuit}, I_{\spadesuit}$. *Radius*$_1$ is valid due to the deductive closure of possible worlds.[10]

Proof for Theorem 3:

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

Definition 10 (Canonical model). *The canonical model for the logic* Λ *is* $\mathcal{M} := \langle \mathcal{W}, \geq, \sim, \mathcal{V}, \mathcal{R}, cp, \mathcal{I}_{\spadesuit}, \hat{\mathcal{I}}_{\spadesuit} \rangle$ *where:*

[10] 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.

- *$\mathcal{W}$ is the set of all maximal Λ-consistent sets.*
- *$\geq$ is such that for $w, u \in \mathcal{W}$: $w{\geq}u$ iff $\{\phi \mid L_{\Box}\phi \in w\} \subseteq u$.*
- *$\sim$ 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 $w \in \mathcal{W}$.*
- *$\mathcal{I}_{\spadesuit}(w) = \{\phi \mid I_{\spadesuit}\phi \in w\}$, with $w \in \mathcal{W}$.*
- *$\hat{\mathcal{I}}_{\spadesuit}(w) = \{\phi \mid \hat{I}_{\spadesuit}\phi \in w\}$, with $w \in \mathcal{W}$.*

There are alternative but equivalent definitions of $\geq$ and $\sim$ in terms of the duals $\hat{L}_{\Box}$ and $\hat{L}_K$, i.e. $\hat{L}_{\Box}\phi := \neg L_{\Box}\neg\phi$ and $\hat{L}_K\phi := \neg L_K\neg\phi$. Then $w{\geq}u$ iff $\{\hat{L}_{\Box}\phi \mid \phi \in u\} \subseteq w$. The existence lemma is obtained by the traditional routine. That is, for any $w \in \mathcal{W}$, if $\hat{L}_{\Box}\phi \in w$ then there is some $v \in \mathcal{W}$ such that $w{\geq}v$ and $\phi \in v$. Analogous claims can be made for $\sim$ and $\hat{L}_K$. Furthermore, due to L_K, $L_{\Box}$, *Indefeasibility*, *Local Connectedness* and modal logic results on correspondence ([8]) 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.

We then perform induction on the complexity of ϕ to show our 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 $\langle RAD\rangle_{\rho}$ follows by the I.H. and *Radius*$_1$. The claims for $\hat{L}_{\Box}$ and $\hat{L}_K$ follow with the help of the existence lemmas and I.H., while for $I_{\Box}$, I_K, $\hat{I}_{\Box}$ and $\hat{I}_K$ we rely on the construction of the awareness-like functions and then the result is immediate. For $K\phi$ and $\Box\phi$, we make use of *Red*$_K$ and *Red*$_{\Box}$, the I.H. and the results of the previous steps on L_K, I_K and $L_{\Box}$, $I_{\Box}$. [11]

Proof for Theorem 4:

Proof.

- The claim is easy for the atoms, the boolean cases, the inequalities, A, $\langle RAD\rangle_{\rho}$, L_K and $L_{\Box}$. We will only show why the claim holds for I_K and $I_{\Box}$, $\hat{I}_K$ and $\hat{I}_{\Box}$ because the claims involving K, $\Box$ 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, $M, w \models I_K[RAD]_{\rho}\phi$ and by $M, w \models (cp \geq c_{\rho})$, $M, w \models A\rho$, we finally

[11] In fact, we can claim that this logic is weakly complete with respect to ALPS where $\geq$ is conversely well-founded. This is because our structures have the finite model property (via filtration theorem, [8]) so there are no infinite $>$ chains of more and more plausible worlds.

get $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_K[RAD]_\rho\phi$. For the other direction, suppose that $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_K[RAD]_\rho\phi$. 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 $M, w \models I_K[RAD]_\rho\phi$, for all $u \in W^I$, $M, u \models [RAD]_\rho\phi$, 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 $M, w \models \langle\rho\rangle I_\Box\phi$. Therefore $M, w \models (cp \geq c_\rho)$, $M, w \models A\rho$ and $M^\rho, w \models I_\Box\phi$. 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 $u \in W^I \cap Q_\Box(w)$ and arbitrary $v \in u^\rho$. Since $ord^\rho(v) \leq ord(u)$ (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 $M, w \models I_\Box[RAD]_\rho\phi$ and by $M, w \models (cp \geq c_\rho)$, $M, w \models A\rho$, we finally get $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_\Box[RAD]_\rho\phi$. For the other direction, suppose that $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I_\Box[RAD]_\rho\phi$. 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 $ord^\rho(v) = ord(u)$ and $w \geq^\rho v$ then $w \geq u$. By the truth conditions of $M, w \models I_\Box[RAD]_\rho\phi$, for all $u \in W^I$ such that $w \geq u$: $M, u \models [RAD]_\rho\phi$, 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, $M^\rho, v \models \phi$ too. Overall, $M^\rho, w \models I_\Box\phi$ and finally $M, w \models \langle\rho\rangle I_\Box\phi$.

- 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 $M, w \models \langle\rho\rangle \hat{I}_\Box\phi$. Therefore $M, w \models (cp \geq c_\rho)$, $M, w \models A\rho$ and $M^\rho, w \models \hat{I}_\Box\phi$. 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 $u \in W^I \cap Q_\Box(w)$ and $v \in u^\rho$ with $M, v \models \phi$, which is precisely $M, w \models \hat{I}_\Box\langle RAD\rangle_\rho\phi$. Overall: $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge \hat{I}_\Box\langle RAD\rangle_\rho\phi$. For the other direction, suppose

$M, w \models (cp \geq c_\rho)$, $M, w \models A\rho \wedge \hat{I}_\Box \langle RAD \rangle_\rho \phi$. This means that there is some $u \in W^I \cap Q_\Box(w)$ and some $v \in u^\rho$ with $M, v \models \phi$. It suffices to show that $M^\rho, w \models \hat{I}_\Box \phi$. 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 $M^\rho, w \models \hat{I}_\Box \phi$. Overall indeed $M, w \models \langle \rho \rangle \hat{I}_\Box \phi$.

References

[1] Alechina, N., Logan, B.: A logic of situated resource-bounded agents. Journal of Logic, Language and Information 18, 79–95 (2009)

[2] Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. In: Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge. pp. 43–56. TARK '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998)

[3] Baltag, A., Renne, B.: Dynamic epistemic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edn. (2016)

[4] Baltag, A., Smets, S.: 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 (2008)

[5] van Benthem, J.: Dynamic logic for belief revision. Journal of Applied Non-Classical Logics 17(2), 129–155 (2007)

[6] van Benthem, J.: Tell it like it is: Information flow in logic. Journal of Peking University (Humanities and Social Science Edition) (2008)

[7] van Benthem, J.: Logical Dynamics of Information and Interaction. Cambridge University Press (2011)

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

[9] Cherniak, C.: Minimal Rationality. Bradford book, MIT Press (1986)

[10] Cowan, N.: The magical number 4 in short-term memory: A reconsideration of mental storage capacity. The Behavioral and Brain Sciences 24, 87–114 (2001)

[11] van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Springer Publishing Company, Incorporated, 1st edn. (2007)

[12] Duc, H.N.: Reasoning about rational, but not logically omniscient, agents. Journal of Logic and Computation 7(5), 633 (1997)

[13] Fagin, R., Halpern, J.Y.: Belief, awareness, and limited reasoning. Artificial Intelligence 34(1), 39–76 (1987)

[14] Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning About Knowledge. MIT Press (1995)

[15] Fagin, R., Halpern, J.Y.: Reasoning about knowledge and probability. Journal of the ACM 41(2), 340–367 (1994)

[16] Hintikka, J.: Impossible possible worlds vindicated. Journal of Philosophical Logic 4(4), 475–484 (1975)

[17] Johnson-Laird, P.N., Byrne, R.M., Schaeken, W.: Propositional reasoning by model. Psychological Review 99(3), 418–439 (1992)

[18] Kahneman, D., Beatty, J.: Pupillary responses in a pitch-discrimination task. Perception & Psychophysics 2(3), 101–105 (1967)

[19] Lehrer, K.: Theory of Knowledge. Westview Press (2000)

[20] Miller, G.: The magical number seven, plus or minus 2: Some limits on our capacity for processing information. Psychological Review 63, 81–97 (1956)

[21] Parikh, R.: 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., Amsterdam, The Netherlands (1987)

[22] Plaza, J.: Logics of public communications. Synthese 158(2), 165–179 (2007)

[23] R Sears, C., Pylyshyn, Z.: Multiple object tracking and attentional processing. Canadian Journal of Experimental Psychology 54, 1–14 (2000)

[24] Rantala, V.: Impossible worlds semantics and logical omniscience. Acta Philosophica Fennica 35, 106–115 (1982)

[25] Rasmussen, M.S.: Dynamic epistemic logic and logical omniscience. Logic and Logical Philosophy 24, 377–399 (2015)

[26] Rasmussen, M.S., Bjerring, J.C.: A dynamic solution to the problem of logical omniscience. Journal of Philosophical Logic (forthcoming)

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

[28] Schroyens, W., Schaeken, W.: A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning 29, 140–9 (2003)

[29] Schroyens, W.J., Schaeken, W., D'Ydewalle, G.: 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 (2001)

[30] Spohn, W.: Ordinal conditional functions. a dynamic theory of epistemic states. In: Harper, W.L., Skyrms, B. (eds.) Causation in Decision, Belief Change, and Statistics, vol. II. Kluwer Academic Publishers (1988)

[31] Stalnaker, R.: On logics of knowledge and belief. Philosophical Studies 128(1), 169–199 (2006)

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

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

[34] Velázquez-Quesada, F.R.: Small Steps in Dynamics of Information. Ph.D. thesis, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands (2011)

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

[36] Wassermann, R.: Resource bounded belief revision. Erkenntnis 50(2), 429–446 (1999)

[37] Xu, Y., Chun, M.M.: Selecting and perceiving multiple visual objects. Trends in Cognitive Sciences 13(4), 167–174 (2009)