Logical Models for Bounded Reasoners

Anthia Solaki

Logical Models for Bounded Reasoners

ILLC Dissertation Series DS-2021-07

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Paris Ioannou.
Printed and bound by Ipskamp Drukkers.

ISBN: 978-94-6421-318-8

Logical Models for Bounded Reasoners

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen
op vrijdag 18 juni 2021, te 14.00 uur

door

Anthia Solaki

geboren te Thessaloniki

Promotiecommisie

Promotores:	Prof.dr. S.J.L. Smets	Universiteit van Amsterdam
	Prof.dr. F. Berto	Universiteit van Amsterdam
Overige leden:	Prof. dr. J.F.A.K. van Benthem	Universiteit van Amsterdam
	Prof. dr. A. Betti	Universiteit van Amsterdam
	Prof. dr. F. Liu	Universiteit van Amsterdam
	Dr. J.K. Szymanik	Universiteit van Amsterdam
	Prof. dr. L.C. Verbrugge	Rijksuniversiteit Groningen
	Prof. dr. L.S. Moss	Indiana University Bloomington
	Prof. dr. P. Egré	École Normale Supérieure

Faculteit der Geesteswetenschappen

The research for this doctoral thesis received financial assistance from the Netherlands Organization for Scientific Research (NWO), grant No. 322-20-018.

Στην οικογένειά μου

Contents

Part I: Motivation

Part II: Single-agent Reasoning

Part III: Multi-agent Reasoning

Acknowledgments

I am indebted to my supervisors, Sonja Smets and Franz Berto. This project would not have started without them putting faith in that long grant application and would not have been completed without their constant support ever since. Thank you, Sonja, for always encouraging me to pursue my ideas and suggesting ways to take them even further; for your advice, not only on the contents of the dissertation, but on academic life in general; for everything I learned by working next to you, in research and in teaching. Thank you, Franz, for believing in me and this project probably before I did; for your enthusiasm and guidance over the years; for welcoming me into the LoC group; and for being the type of supervisor who understands that a concert of Roger Waters is a valid reason to leave an event early in order to catch a plane back to Amsterdam. Thank you both for being such genuinely nice people, with whom I could always talk openly.

I am also grateful to Johan van Benthem, Arianna Betti, Fenrong Liu, Jakub Szymanik, Rineke Verbrugge, Larry Moss, and Paul Egré, for agreeing to be in my doctorate committee and for the time they put in reading this dissertation.

The dissertation has benefited immensely from my exchanges with brilliant researchers in Amsterdam, in Paris, and in other corners of the (lately, virtual) world, where I was given the opportunity to present and discuss my work. Special thanks to Alexandru Baltag, Johan van Benthem, Paul Egré, Dan Hoek, the members of the Milan Logic group, the audiences of the "Logic in the Wild" workshop, Logica 2018, WoLLIC 2018, LORI 2019, DaLí 2020. Their remarks have substantially improved my results. Still, this list would be much longer if I mentioned everyone who has contributed to this process.

That being said, I cannot but make a special reference to my academic home over the past years, the ILLC. Thanks to the ILLC staff, whose support has been essential, especially amidst the difficult circumstances of the last year. Thanks to my PhD colleagues and friends, whose wit and open mind make even a short conversation fun and inspiring. Thanks to everyone who has been involved in the LIRa seminar for making it what it is: an intellectually stimulating environment where scientific exchange and inspiration flourishes, while still leaving plenty of room for good laughs and amusing sociopolitical commentary. Thanks to the LoC group which has also felt like an academic family to me. Thank you, Aybüke

Özgün, Karolina Krzyżanowska, Peter Hawke, and Tom Schoonen, for our very fruitful and cozy meetings. I have benefited a lot from your constructive comments and from the way you work, individually and collectively.

I would like to express my gratitude to Fernando Velázquez-Quesada. Thank you, Fer, for being a great researcher and collaborator whose work and approach to problem-solving have shaped my own; for all the feedback you provided, including on parts of this dissertation; for the interesting discussions over topics beyond research, and for the pleasant memories of climbing rocks in Czechia and fences in Bulgaria (invaluable life skills). To my ILLC office-mates over the years: thank you, Ilaria Canavotto, Gianluca Grilletti, Tom Schoonen, Chenwei Shi, Thom van Gessel, Kaibo Xie, for creating a welcoming environment to share work and non-work related curiosities and frustrations as well as delicacies from all around the world (I will explicitly not count *zoute drop* here; sorry, guys). Thank you for your friendship and for those refreshing drinks and barbecues. This cannot but also apply to my OT office-mate and friend Aybüke, who apart from being a bright logician whose opinion is always valuable, is also such a kind person – in fact, that kind to deal with my (bad) jokes. Extra thanks to Ilaria and Tom for patiently answering my many questions on the details of writing a dissertation, reading parts of it, providing very helpful feedback, and listening to my worries. Thanks to Tom, for his help in writing the samenvatting too.

I wouldn't look back on this adventure with a smile, if it weren't for some special people I'm fortunate to be surrounded by in the Netherlands and in Greece. Vera, you have a way of making everything look easier, even when the problems (real or made-up by me) seem insurmountable. Your presence makes the cliché phrase "*komt goed*" actually helpful. Bedankt voor *alles*. José, Jo, Wim, en Wendy, bedankt voor jullie vriendelijkheid en de gezellige avondjes. Thank you, Anastasia, Dina, Marianna, and Vicky for your uncanny ability of knowing when to be serious, and perhaps more importantly, knowing when not to. I can't imagine the last many years without your unfailing support and contagious laughter.

My deepest gratitude goes to my family. Mom and dad, I know you probably won't read any of the pages that follow. But I also know that without your love, your support, your faith, and your example, they would not have been written at all. Ευχαριστώ. All these years, I could count on my sister no matter what. Thank you for making me confident that I won't encounter the *Laestrygonians, the Cyclops, the wild Poseidon* of our favourite poem. Thano, you used to keep me company the nights when I kept the light on to study while you probably just wanted to sleep. That annoying trait of mine was there to stay, but so was your patience, for which I am extremely grateful. And thank you, Marilia and Danae, for your smiles, for the works of art you kindly donated, and for making sure my geography knowledge never gets rusty.

Anthia Solaki
Amsterdam, April 2021

Chapter 1

Outline

This thesis is broadly concerned with the logical modelling of aspects of human reasoning, informed by facts on the bounds of human cognition. Through this investigation, we aim at building more bridges between logic and disciplines studying human reasoning from an empirical perspective. We will pursue this overarching goal by breaking it down into three parts, each dealing with a particular family of questions arising in the process.

We lay the foundations for this attempt in Part I. The central question we seek to answer is, simply put: *Why should we design logical systems for bounded reasoners?*

The answer is given by Chapter 2. We first present the received view of *Epistemic Logic*, which is occupied with the formal study of knowledge and belief, and some of the philosophical objections against it. We then survey empirical evidence on human reasoning, that seems to corroborate objections against the adequacy of the received view in modelling aspects of human reasoning. In particular, empirical findings reveal the limits of real reasoners when it comes to (i) deductive reasoning (e.g. we do not know all logical consequences of our knowledge), (ii) introspection (e.g. we do not always have infallible access to our own beliefs), and (iii) reasoning about others (e.g. we do not always reason correctly about what our friend believes about what we believe). The contrast between standard Epistemic Logic and empirical evidence is embedded in the so-called *Rationality Debate*, which is occupied with questions such as: "*Are humans rational? If so, in what sense?*". We investigate whether the traditional rationality norms, to which logical formalizations often correspond, can and should be defended in light of the empirical evidence. After some critical discussion, we identify features of an alternative picture of rationality, and propose one wherein descriptive facts have a key role. As a result, we argue that logical systems should also be revised, in order to stand as the formal counterparts of the alternative view. We identify several sub-tasks of this project, which are pursued in the next chapters.

In Part II, we set out to address these tasks, starting from those concerning the question: *How can we model the reasoning endeavours of a single agent?* Chapters 3 to 5, guided by the desiderata of Chapter 2, take seriously into account the resource-boundedness of human reasoners and the variety of processes underlying their reasoning.

In Chapter 3, we design a logical framework for the deductive reasoning of a single bounded agent. The main contribution is the introduction of a *resource-sensitive* syntax and semantics, which make use of *Dynamic Epistemic Logic* and *impossible-worlds semantics*. The core idea is to explicitly encode the agent's deductive reasoning steps in the logical system and to monitor the cognitive effort they require. As a result, we account for a fallible yet logically competent agent, who performs reasoning processes, but only to the extent allowed by cognitive resources. We discuss how this framework fulfils some of the desiderata of Chapter 2 and examine its technical features. In particular, we show that there is a connection between resource-sensitive models and *syntactic structures*, which allow for a detour towards extracting a sound and complete axiomatization. Apart from addressing an aspect of the envisaged alternative rationality picture, Chapter 3 also serves as the building block for the richer frameworks of the next chapters.

In Chapter 4, we combine the basic resource-sensitive framework with *plausibility models*, which have been used in Dynamic Epistemic Logic to account for a greater repertoire of notions of knowledge and belief, and policies of integrating incoming information. Therefore, this combination brings us closer to the nuances of a human agent's mental states. The technical contribution also extends to the plausibility modelling, by establishing connections between different types of structures and manifesting their added value in obtaining a sound and complete axiomatization. Another benefit of our plausibility framework is that it facilitates the study of the mixed nature of human reasoning, which involves both steps of "internal" deductive reasoning and actions of "external" information, provided by sources of varying reliability.

In Chapter 5, we focus on another component of the alternative picture, arising from empirical evidence on human reasoning. In particular, we aim at modelling the logical aspects of a distinction of mental processes that has played a key-role in revisiting the understanding of rationality in many disciplines, namely the distinction between *fast and slow thinking*. The plausibility modelling of Chapter 4 is instrumental in this investigation. The resulting framework does justice to the workings of the different processes as well as to their interactions. It can therefore encompass the study of reasoning scenarios that are often neglected in the logical literature. In this way, we realize another connection between logical formalizations and elements of the alternative picture. On basis of this, we argue for the normative status of our logical models, which, contrary to the ones of the received view, seek normative standards that are actually attainable by bounded reasoners.

In Part III, we move from single- to multi-agent reasoning. The central question is: *How can we model bounded reasoning in a multi-agent environment?* This reveals additional layers of investigation compared to Part II, involving, for example, *higher-order reasoning* and *group reasoning*. In accord with the division of labour fleshed out in Chapter 2, we unfold this investigation in three directions: (a) the formation of beliefs about others, (b) the manipulation of formed beliefs in a multi-agent environment, and (c) the effect of the former on group reasoning.

In Chapter 6, we focus on direction (a). We first analyze which elements of the semantics of Epistemic Logic might make it less suitable for the modelling of the agents' higher-order reasoning. On basis of this analysis, we propose a temporal framework for modelling mental state attributions that are formed by observation, memory, and communication. Due to our interest in building bridges between logic and empirical studies, we apply this framework to *False Belief Tasks*, a well-known family of experiments that test the ability of people to attribute mental states to others. We also investigate the technical properties of the setting and its relationship to well-known modal temporal logics. We evaluate our results both with respect to desiderata pertaining to the formalization of the tasks, and with respect to cognitive features pertaining to the alternative picture.

In Chapter 7, we focus on direction (b). We study the manipulation of beliefs in a multi-agent context, not only via acts of deductive reasoning, but also via acts of reasoning about oneself and others. This is achieved by the design of multi-agent resource-sensitive models and of special *action models*, which are compatible with impossible-worlds semantics and able to represent uniformly the reasoning steps of deductive inference, introspection, and reasoning about others. We show that these steps, when cognitively affordable, can refine the zero- and higher-order beliefs of agents. We also expand our method for the extraction of a sound and complete axiomatization, by establishing a correspondence between our models and multi-agent syntactic structures.

In Chapter 8, we focus on direction (c). The objections against Epistemic Logic are often inherited by its formalizations of group epistemic notions. Our case in point is the notion of *distributed knowledge*, which is taken to represent what *would* be known, *if* group members were to pool their knowledge and infer information on its basis. This traditional understanding of distributed knowledge neglects the fact that the cognitive capacities of group members interfere with what can be actually achieved by the group. Inspired by experiments on group reasoning, we identify two dimensions that shed light on whether and how a group "actualizes" its distributed knowledge, namely *communication* and *inference*. We build a dynamic framework with effortful actions accounting for both, using our multi-agent resource-sensitive semantics. On the more technical side, we bring together the methods of Chapter 7 with common logics of distributed knowledge.

We conclude with Chapter 9, where we briefly discuss two general directions for future work and we reflect on the contributions of the thesis and as a whole.

Origin of the material

The thesis is based on the articles listed below. The articles are (co-)authored by Anthia Solaki; co-authors (if any) and their relative contributions are specified for each item. The content of the articles has been adapted to increase the cohesion of the thesis. We indicate the origins of each chapter on basis of the listed items.[1]

[1] Where is Epistemic Logic in the Rationality Debate? [submitted manuscript]

[2] A dynamic epistemic logic for resource-bounded agents. *The Logica Yearbook 2018.* pages 229-244. College Publications. 2019.

[3] The effort of reasoning: Modelling the inference steps of boundedly rational agents. In *International Workshop on Logic, Language, Information, and Computation*, pages 307–324. Springer. 2018. [with Sonja Smets]

> Both authors discussed the central ideas of the article together. In the writing stage, Sonja Smets developed the introduction and Anthia Solaki developed the main framework and its applications.

[4] The effort of reasoning: Modelling the inference steps of boundedly rational agents. *Journal of Logic, Language and Information.* [to appear, extended version of the previous item, with Sonja Smets, authors' contributions as above]

[5] The logic of fast and slow thinking. *Erkenntnis*, pages 1–30. 2019. [with Francesco Berto and Sonja Smets]

> The authors discussed the central ideas of the article together. In the writing stage, the philosophical background and conclusions were developed by Francesco Berto and the logical background was developed by Sonja Smets. Anthia Solaki developed the main framework and its applications.

[6] Towards a logical formalisation of Theory of Mind: A study on false belief tasks. In *International Workshop on Logic, Rationality and Interaction*, pages 297–312. Springer. 2019. [with Fernando R. Velázquez-Quesada]

> Both authors contributed equally in all stages.

[7] A logical formalisation of False Belief Tasks. [submitted manuscript, extended version of the previous item, with Fernando R. Velázquez-Quesada, authors' contributions as above]

[1] The following article was also written during the PhD project, but it does not correspond directly to a chapter of the thesis: Rule-based reasoners in epistemic logic. In *European Summer School in Logic, Language and Information*, pages 144-156. Springer. 2018.

[8] Bounded multi-agent reasoning: Inference, introspection, attribution. [submitted manuscript]

[9] Bounded multi-agent reasoning: Actualizing distributed knowledge. In *International Workshop on Dynamic Logic*, pages 239–258. Springer. 2020.

[10] Actualizing distributed knowledge in bounded groups. [submitted manuscript, extended version of the previous item]

[11] What do you believe your friends believe? Towards realistic belief attributions in multi-agent systems. In *NETREASON workshop at 24th European Conference on Artificial Intelligence*. 2020. [extended abstract, with Fernando R. Velázquez-Quesada, authors' contributions as above]

Chapter 2 is mainly based on [1]
Chapter 3 is mainly based on [2,3,4]
Chapter 4 is mainly based on [3,4,2]
Chapter 5 is mainly based on [5]

Chapter 6 is mainly based on [6,7]
Chapter 7 is mainly based on [8]
Chapter 8 is mainly based on [9,10]
Chapter 9 is partly based on [11]

Part I

Motivation

Chapter 2

Where is Epistemic Logic in the Rationality Debate?

In this chapter, we explain the motivation behind the construction of logical models for bounded reasoners.[1] We discuss the place of logical systems for knowledge and belief in the *Rationality Debate*, i.e. the debate on whether, and in what sense, humans are rational or not. In particular, we present the standard paradigm of Epistemic Logic, juxtaposed with empirical work on human reasoning. We then consider the philosophical implications of this and identify features of an alternative picture of rationality. As a result, we argue that logical systems should be revised in order to stand as formal counterparts of the alternative rationality view.

2.1 Introduction

Epistemic Logic (EL) is concerned with the formal study of knowledge and other propositional attitudes, like belief. The added value of a formal study lies in revealing systematic features of epistemic notions and advancing the debate on open epistemological problems. However, EL, at least in its conventional understanding as a spinoff of Modal Logic, has also been criticized on grounds of its adequacy to study the epistemic endeavours of real people. There have been objections raised against its idealized modelling of our activities: the way we perform deduction, reflect upon our knowledge, or ascribe mental states to others. The models of EL are usually defended by appealing to the need for simplicity, even at the cost of accuracy, or by arguing that their purpose is to be normative, not descriptive of people's behaviour: they only predict how people *ought to reason*.

However, much of this debate has remained strictly within the circles of logicians, philosophers, and computer scientists. Contrary to that, we find it important to zoom out and consider interdisciplinary empirical evidence on human reasoning. Epistemology, broadly, and EL, in particular, are concerned with hu-

[1] The chapter is based on Solaki (2021c).

man reasoning. It only makes sense to seriously take into account how people *actually* reason, a topic studied by psychologists and cognitive scientists, among others. This does not go against the normative purposes of epistemology or logical formalizations thereof. Setting appropriate principles that people ought to follow should at least be properly informed by what people actually do when they reason.

More specifically, empirical evidence seems to threaten a view on rationality that is assumed in traditional epistemology and to which the agents of EL correspond. This view roughly dictates that people are rational animals, complying with the rules of (classical) logic. But when empirical evidence enters the picture, this claim is on shaky ground. This has led to a lively debate, the Rationality Debate. The interpretations of experimental data by different theorists have implications for the traditional view of rationality: from downright rejecting it to suggesting more or less radical modifications of it. The implications for epistemology then extend to the logical formalizations as well.

In this chapter we discuss the place of logical systems for knowledge and belief in the Rationality Debate. We first present the standard paradigm of EL, using Kripke semantics, and present some of the philosophical problems and objections against it (Section 2.2). We then survey empirical work on human reasoning, discussing the results of influential reasoning tasks and the analyses in which these findings have been embedded due to their implications for the Rationality Debate (Section 2.3).

In Section 2.4, we go deeper in interpreting these results and their significance in settling the Debate. We investigate whether the traditional understanding of rationality, as compliance with (classical) logic, can and should be defended in light of the psychological evidence. After some critical discussion of arguments put forward in this context, we identify some features of an alternative picture of rationality, and propose one wherein descriptive facts have a key role. Once we have discussed the need to shift to a more naturalized notion of rationality, we argue that the logical systems should also be revised, in order to stand as the formal counterparts of this alternative rationality view (Section 2.5).

2.2 Epistemic Logic

2.2.1 The received view

The work of von Wright (1951) and Hintikka (1962) has laid the foundations for the application of Modal Logic to the formal study of propositional attitudes such as knowledge and belief. Standard epistemic and doxastic logics were developed as spinoffs of Modal Logic, making use of *relational possible-worlds semantics*. The core idea is that in knowing or believing something, one obtains a way of determining which the *actual* world is among a range of possibilities (see "information as range" in van Benthem et al. (2008)). *Possible worlds* embody precisely this conception of logical possibilities.[2]

The standard approach accounts for knowledge by supplementing the language of propositional logic with a unary operator K such that $K\phi$ reads: "the agent knows that ϕ". Following the same fashion, we can add a unary operator B such that $B\phi$ reads "the agent believes that ϕ". Next, the semantic interpretations are given in terms of possible worlds: an agent knows(/believes) that ϕ if and only if in all possible worlds that are epistemically(/doxastically) accessible to her, it is the case that ϕ. There can be more than one operator to accommodate settings with more than one agent. Then, by indexing the operators, we get $K_i\phi$, read as "agent i knows that ϕ", and likewise for belief. What follows can be easily generalized for multi-agent settings.

We will give the concrete account of standard single-agent epistemic logic, starting off with the constructions of Modal Logic, and also comment on how this can be adapted for doxastic and combined epistemic-doxastic frameworks.

2.2.1. Definition (Language). The language of single-agent epistemic logic is defined as follows:

$$\phi \quad ::= \quad p \quad | \quad \neg\phi \quad | \quad \phi \wedge \phi \quad | \quad K\phi$$

with $p \in \Phi$ and Φ a given set of propositional atoms.

The language of single-agent epistemic-doxastic logic is easily obtained by supplementing the previous definition with $B\phi$. The common Boolean connectives are defined in terms of $\neg$ and $\wedge$.

Next, we show how the relational structures of Modal Logic are utilized in the epistemic context. More specifically, the compatibility of worlds with the agent's knowledge (or belief) is captured via primitive binary relations on possible worlds.

2.2.2. Definition (Kripke frames and models).

1. A *Kripke frame* is a pair $\mathcal{F} = \langle W, R \rangle$, where W is a non-empty set of possible worlds and R is a binary *accessibility* relation on W.
2. A *Kripke model* is a frame supplemented with a valuation $V : W \to \mathcal{P}(\Phi)$ assigning to each $w \in W$ a subset $V(w)$ of Φ. Intuitively, $V(w)$ is the set of all propositional atoms that are true at w.

The accessibility relation can be used to denote epistemic or doxastic accessibility. Frames and models might be endowed with more than one accessibility relation, thereby allowing for combined epistemic-doxastic settings (and multi-agent settings too). We proceed with the truth clauses and other key-definitions:

[2] Detailed explanations, along with the historical background behind the emergence of the field, are given by Rendsvig and Symons (2019). The overview presented in this section (definitions, terminology, etc.) is based on standard sources for Modal Logic and EL (Fagin et al., 1995; Blackburn et al., 2001; Bezhanishvili and Van Der Hoek, 2014) and draws on material from the author's master's thesis (Solaki, 2017).

2.2.3. Definition (Truth). For a world w in a model $M = \langle W, R, V\rangle$, we define that a formula ϕ is *true in M at world w* (notation: $M, w \models \phi$) as follows:

$$\begin{array}{lll}
M, w \models p & \text{iff} & p \in V(w), \text{ where } p \in \Phi \\
M, w \models \neg\phi & \text{iff} & M, w \not\models \phi \\
M, w \models \phi \wedge \psi & \text{iff} & M, w \models \phi \text{ and } M, w \models \psi \\
M, w \models K\phi & \text{iff} & \text{for all worlds } u \in W \text{ such that } wRu\text{: } M, u \models \phi
\end{array}$$

Regarding belief, and denoting the doxastic relation with R_b, we can simply define an extended frame $\mathcal{F} = \langle W, R, R_b\rangle$ and model $M = \langle W, R, R_b, V\rangle$ as suggested above. Then the semantic interpretation for belief is given by:

$$M, w \models B\phi \qquad \text{iff} \qquad \text{for all worlds } u \in W \text{ such that } wR_bu\text{: } M, u \models \phi$$

2.2.4. Definition (Truth in a model). A formula is *true* (or *valid*) *in a model* if it is true at all possible worlds of the model.

2.2.5. Definition (Frame validity). A formula is *valid at a world w in a frame $\mathcal{F}$* if it is true at w in every model $\langle \mathcal{F}, V\rangle$ based on $\mathcal{F}$. It is *valid in a frame $\mathcal{F}$* if it is valid at every world w in $\mathcal{F}$. It is *valid in a class of frames* if it is valid in every frame of the class.

The use of *correspondence results* (Blackburn et al., 2001) renders many properties of knowledge and belief amenable to formal study. In particular, the validity of certain formulas is associated with certain properties of the accessibility relation(s). The following definition is useful in investigating this correspondence:

2.2.6. Definition (Normal modal epistemic logic). A *normal modal epistemic logic* Λ is a set of formulas that contains all instances of propositional tautologies, all instances of the *Kripke schema* (K): $K(\phi \to \psi) \to (K\phi \to K\psi)$ and is closed under *Modus Ponens* and the *Necessitation Rule* (N): from ϕ infer $K\phi$.

The normal doxastic logic is obtained along these lines. Certain formulas characterize classes of frames whose accessibility relations have certain algebraic properties. The classes of frames determined by those properties reflect useful properties of knowledge and belief, often revealing connections with epistemological corollaries. To begin with, the class of all frames corresponds to the smallest normal modal logic, which is called **K**. Extensions of this logic are obtained by adding axioms that seem plausible according to our intuitive understanding of knowledge/belief and the epistemological discussion that has long investigated how these attitudes can be discerned. We give an overview of properties that have been suggested for the adequate formal description of knowledge and belief.

Veridicality. The axiom scheme that reflects the *veridicality of knowledge*, i.e. that if ϕ is known then ϕ is true, is called T: $K\phi \to \phi$. Adding T to the logic **K**

results in the logic **T**. The axiom T corresponds to the class of those frames where for every world w, wRw, i.e. the class of *reflexive* frames. Likewise, if one accepts veridicality of belief, the belief-version of the scheme should be added, turning R_b into a reflexive relation too. Veridicality for belief is usually not assumed.

Consistency. The axiom scheme that reflects the *consistency of knowledge* is called D: $K\phi \to \neg K\neg\phi$ (or equivalently: $\neg K(\phi \wedge \neg\phi)$). Adding D to the logic **K** results in the logic **D**. The axiom is valid exactly on those frames where for any world w, there is some world u such that wRu, i.e. the class of all *serial* frames. Accordingly, the belief version of the axiom is $B\phi \to \neg B\neg\phi$ and corresponds to seriality of the doxastic accessibility relation.

Positive Introspection. The axiom scheme (4): $K\phi \to KK\phi$ reflects the *positive introspection of knowledge*. The addition of (4) to the logic **K** yields the logic **K4**. It characterizes the class of those frames where for any worlds w, u, v, if wRu and uRv then wRv, i.e. the class of all *transitive* frames. Positive introspection of belief works along the same lines.

Negative Introspection. The axiom scheme (5): $\neg K\phi \to K\neg K\phi$ reflects the *negative introspection of knowledge* and adding it to **K** results in the logic **K5**. This axiom scheme characterizes the class of those frames where for any worlds w, u, v, if wRu and wRv then uRv, i.e. the class of all *euclidean* frames. Negative introspection of belief again works along these lines.

Veridicality is often seen as an essential property of knowledge, but not of belief. In fact, it has been argued that this is one of the properties that can be used to distinguish knowledge and belief. On the other hand, the appeal of positive and negative introspection may be debated for both knowledge and belief, as is consistency of belief.

Overall, combinations of these axioms result in logical systems of varying strength that are sound and complete with respect to those classes of frames with the corresponding properties of the accessibility relation(s). Picking the *most* appropriate system depends on one's goals and preferred analysis of propositional attitudes. Still, according to the received view (e.g. as in Fagin et al. (1995); van Ditmarsch et al. (2007)) (a) epistemic models are S5-models, i.e. models in which the (epistemic) accessibility relation is an equivalence relation (reflexive, transitive, and symmetric) and (b) doxastic models are KD45-models, i.e. models in which the (doxastic) accessibility relation is serial, transitive, and euclidean.

Logic	Axioms	Class of frames
K	K	All
KD45	K, D, (4), (5)	Serial, Transitive, Euclidean
S4	K, T, (4)	Reflexive, Transitive
S5	K, T, (5)	Reflexive, Transitive, Symmetric

Table 2.1: Common modal logics

This is a non-exhaustive list. Other axiom choices, that have not been mentioned for reasons of brevity, can yield other, intermediate systems, e.g. between **S4** and **S5**. One may suggest these systems as better candidates for the modelling of knowledge or belief. For simplicity, we sometimes use solely the term "Epistemic Logic" to refer to the received view on both knowledge and belief.

As mentioned above, these systems can be easily extended to multi-agent settings. In such settings, we are interested not only in the individual knowledge or beliefs of each agent, but also in group epistemic notions (Fagin et al., 1995, Chapter 2). For example, *mutual knowledge* of ϕ: everybody knows that ϕ; *common knowledge* of ϕ: everybody knows that ϕ, everybody knows that everybody knows that ϕ, and so on, *ad infinitum*; *distributed knowledge* of ϕ: whenever agents, by pooling their knowledge together, can deduce ϕ. These notions have been included in extended logical systems, building on the foregoing. In this way, logical modelling contributes to epistemological discussions and philosophical puzzles that touch upon the social (i.e. not strictly individualistic) aspect as well.

Notice, however, that these logics are purely *static*: they do not account for *changes* in knowledge and beliefs, nor for actions that lead to their formation. Addressing this is exactly the goal of *Dynamic Epistemic Logic* (DEL), the study of modal logics of model change (Baltag et al., 1998; van Ditmarsch et al., 2007; van Benthem, 2011; Baltag and Renne, 2016). With their additional machinery, they do justice to the *dynamic* processes underlying knowledge acquisition and belief revision. While (D)EL has contributed to core issues of epistemology, its contributions go beyond the philosopher's interests. It has been instrumental in computer science as well (Fagin et al., 1995; van Ditmarsch et al., 2015). Often, properties of knowledge, such as negative introspection, over which there is no consensus among philosophers, are assumed by computer scientists due to their applicability to their subject matter, e.g. to distributed computing.

2.2.2 Objections

Despite the elegance of using normal modal logics in the formal study of knowledge and belief, there is a certain cost that comes with it. In particular, there are objections regarding their adequacy in modelling deductive reasoning, introspection, and higher-order reasoning, especially among some logicians and philosophers. In this subsection, we present the gist of these problems, before delving into the empirical evidence that further substantiates these concerns.

- **Logical Omniscience.**

  The *problem of logical omniscience* (Fagin et al., 1995, Chapter 9) is an inherent defect of the use of Kripke semantics in EL. Possible worlds are logically closed, therefore an agent who knows ϕ, automatically knows all logical consequences of ϕ. Notice that the problem can be easily restated for belief. The predictions of this approach are therefore inaccurate; the brightest mathematician might know all axioms of set theory without thereby knowing all their consequences. As we will see, the performance of real agents is evidently inhibited by their limited memory, time pressure, biases, etc.

Equally alarming considerations arise from special cases of the problem. Even if certain modifications alter the kind of structures and the notion of truth in a manner that avoids the full problem, forms of it are retained through weaker problematic closure principles. Some of these special and weaker forms are given below, following the work of Fagin et al. (1995) and van Ditmarsch et al. (2007).

1. If ϕ is valid, then the agent knows ϕ. (*Knowledge of valid formulas*)
2. If the agent knows ϕ and ϕ is logically equivalent to ψ, then the agent knows ψ. (*Closure under Logical Equivalence*)
3. If the agent knows ϕ and also knows $\phi \to \psi$, then the agent knows ψ. (*Closure under Known Material Implication*)
4. If the agent knows ϕ and also knows ψ, then the agent knows $\phi \wedge \psi$. (*Closure under Conjunction*)

Another counter-intuitive quality that is nevertheless attributed to agents, at least under systems that contain the axiom D, is that of Consistency: agents never know/believe both ϕ and $\neg\phi$. However, in the real-world, cognitively limited reasoners often maintain inconsistent beliefs. Moreover, it has been claimed that agents might even believe an *explicit* contradiction and consider themselves justified in doing so. For example, *dialetheists* believe that a particular sentence, the *liar sentence*, is simultaneously true and false (Priest, 2006).

It is worth noticing that this kind of idealization is also inherited by attempts of logical modelling on belief change, e.g. based on the AGM postulates (Alchourrón et al., 1985), which has also been embedded in DEL systems such as (Baltag and Smets, 2008b).

- **Positive and Negative Introspection.**

  There are also philosophical objections against positive and negative introspection. Depending on one's preferred theory of knowledge, these properties may or may not be fulfilled. For example, it has been suggested that many externalist analyses are incompatible with these properties (Dretske, 2004). Others, such as the *defeasibility analysis* (Stalnaker, 2006), also formalized by Baltag and Smets (2008b), hold a more favourable spot to positive, but not negative, introspection. In the analysis of Lemmon (1967), an agent a knows that ϕ iff a has learned that ϕ and has not forgotten that ϕ. This, along with some minor assumptions on forgetting, invalidates positive introspection. The same goes with the analysis of Danto (1967), who argues that an agent knows something only if the agent understands it. Then positive introspection fails because for the agent to understand that she knows ϕ, she needs to understand the concept of knowledge, i.e. possess an adequate theory of knowledge; this cannot be taken for granted and it is not guaranteed by knowing ϕ alone.

Williamson also argues that a concept of *inexact knowledge*, which is necessary for non-ideal agents, cannot be credulously paired with introspection (Williamson, 1992, 2000). This is part of a general line of argumentation grounded on the *luminosity paradox*. A mental state is said to be *luminous*, roughly, when the occurrence of it entails that one is in a position to know that they are in that state (Williamson, 2000, Chapter 4). According to the author, no states that can be gradually gained or lost, i.e. no non-trivial mental states, are luminous. Since "knowing that ϕ" is such a state, positive introspection must fail. Bonnay and Égré (2009, 2011) argue for pairing inexact knowledge and a more moderate and psychologically plausible view on introspection, that is parameterized by resources.

There are additional arguments targeted especially at negative introspection. For example, Lenzen (1978) argues against it considering cases where people, mistakenly, believe that they know something. The author further proposes that a logic for knowledge should be stronger than **S4** but less strong than **S5**. Williamson sees the failure of negative introspection as a demonstration of limits in self-knowledge, since people cannot survey the totality of their knowledge (Williamson, 2000, p.317). Moreover, assuming (a) an understanding of knowledge in terms of belief (*knowledge implies belief*), (b) consistency of belief, and (c) negative introspection, one ends up endorsing that an agent is certain both of knowing and of not knowing something (Halpern, 1996). Williamson (2001) provides a simple proof that negative introspection, along with similar assumptions, forces us to accept the infallibility of the agents' beliefs: whatever they believe is true. This is clearly unrealistic. Negative introspection is already challenged by Hintikka (1962), exploiting the connection between symmetry – which the author claims should be dropped – and negative introspection in (T)-including logics.

- **Reasoning about others at any modal depth.**

  Moving to multi-agent settings, apart from the objections on deductive reasoning and introspection, we have additional reasons to worry. Real people cannot reason at any depth about others, which is also supported by interdisciplinary evidence. Parikh (2007) explains how the problem of logical omniscience can manifest itself in multi-agent systems as well. On top of the limitations concerning an agent's individual reasoning, there are additional issues concerning the agents' reasoning about one another, e.g. about another's capabilities, that might interfere with the predictions of the standard systems. This is especially important for the study of group notions of knowledge and belief. For example, the infinitary understanding of common knowledge in combination with limitations in time and memory, renders it questionable whether it is actually achieved in real contexts. Common knowledge is often seen as a prerequisite for action, but as Clark and Marshall (1981) observe, finite reasoners cannot possibly check infinite conditions, e.g. when making or interpreting definite reference; instead they might be using a family of heuristics, such as *co-presence heuristics*. In fact, it is argued that real people act on a large, but still finite, degree of mutual knowledge instead (*almost common knowledge* (Rubinstein, 1989)). These studies also reveal subtleties in the understanding of DEL tools, like *public announcements* in realizing common knowledge (that is, in logics that contain such operators) (Baltag and Renne, 2016). Verbrugge (2009) focuses on higher-order social cognition and argues for the importance of accurately modelling how agents reason about one another's mental states for a variety of purposes, similarly identifying the problems of the received view with respect to these. The same worries can be extended to the study of *common belief* and *distributed knowledge*; for example, the latter's "classical" definition presupposes that agents pool their knowledge through possibly unlimited actions of communication and that they can immediately unpack all consequences of it. This is as strong an assumption as in the individual case. Another family of philosophical objections against the standard *reductionist* formalization of collective notions is put forward by Gaudou et al. (2015), who propose an alternative, *non-reductionist* way to formalize group belief. Therefore, the idealized view of reasoning about others seems to be attacked from multiple fronts.

2.3 Empirical Evidence

In this section, we survey some of the empirical evidence that uncovers systematic violations of the idea of perfectly rational agents, formalized by EL. We present this evidence regarding (i) the deductive reasoning of real human subjects, (ii) their introspective abilities, and (iii) their reasoning about others' reasoning. This will allow us to re-view the Rationality Debate and re-consider the properties of the received logical view in depth.

2.3.1 Deductive Reasoning

First, we present evidence on the deductive reasoning of human agents, focusing on some of the most influential results and the lively discussion they have sparked.

Wason Selection Task. Probably the most paradigmatic deductive reasoning task, “the mother of all reasoning tasks” (Stenning and van Lambalgen, 2008), is the Wason Selection Task (Wason, 1966, 1968). Below is one of its variants:

> The participants are presented with four cards. Each card has a number on one side and a letter on the other. The visible sides of the cards read A, K, 4, and 7. The participants are asked which cards need to be turned to check whether the following holds: *if a card has a vowel on one side, it has an even number on the other.*

The vast majority of subjects replied that cards A and 4 should be turned or the card A alone. The correct answer, however, is that cards A and 7 should be turned. Card 4 does not have to be turned, for it is irrelevant in testing the validity of the rule, while card 7 can falsify the rule, in case there is a vowel on the other side. Overall, less than 10% managed to give the correct answer in the early formulations of the task, which required the checking of such an abstract statement (Wason and Johnson-Laird, 1972). However, the task only involved, in logicians’ terms, two rules of inference: *Modus Ponens* and *Modus Tollens*. This result challenges the comfortable view of humans as rational animals, who master the use of classical propositional logic and form knowledge and beliefs on its basis.

In the aftermath of Wason’s results, many variants of the task have been proposed and tested, giving rise to a broader debate on what the underlying mechanism of reasoning is.

Selection task variants. The original task, despite its simplicity, is proven to be cognitively hard. Researchers have investigated whether the *contents* of the rule of the task make it easier or difficult. According to Johnson-Laird et al. (1972), *familiarity* with the rule is of key importance since subjects performed much better when the rule was given as a familiar postal regulation (“if a letter has a second class stamp, it is left unsealed”). Griggs and Cox (1982) have similarly shown that a rule involving a social regulation, in particular a drinking law (“if you drink alcohol here, you have to be over 18”) produced good performance as well. Griggs (1995) further investigated whether tweaking other parameters such as instruction, sentence clarification, and decision justification improved performance, observing a facilitation effect when these were combined. According to Cosmides (1989) and Cosmides and Tooby (1992), it is social contract rules that produce good performance because humans have developed cheater detection algorithms in the course of evolution. Still, Isaac et al. (2014) observe that even under this view, logic plays a role, albeit hardwired in a domain-specific module. Another explanation for the poor performance in the abstract variants is that participants do not use logical reasoning, but a heuristic called *matching bias* instead (Evans and Lynch, 1973; Evans et al., 1999). In this view, subjects see information that matches the content of the rule as relevant (cards A and 4) and information that fails to match as irrelevant (card 7).

These results seem to threaten logic's role in human reasoning, in that they reveal a gap between content-dependent rules, that apparently humans actually use, and the content-independent rules of formal (classical) logic. Still, it might be possible to mediate the conflict through the *memory queuing hypothesis* (Griggs and Cox, 1982). According to it, performance is facilitated when the content of the task is familiar to the subjects and allows them to recall past experience. This is compatible with the idea that human reasoning competence is characterized by content-independent rules, which can be applied provided that a situation is *familiar*. The diagnosis of the problem with the *abstract* task is that it does not trigger the appropriate reasoning principle. However, experimental work challenges this hypothesis (Cosmides, 1989). The same rule elicited different responses, when posed as part of an unfamiliar social contract, and when posed as part of an unfamiliar descriptive story. Subjects of the former did much better than the ones of the latter. If the memory queuing hypothesis was correct, the subjects would do equally bad in both variants since the situation was unfamiliar. In (Manktelow and Evans, 1979), subjects performed almost as poorly as in the original task when given a rule about the eating and drinking habits of the experimenter, regardless of the familiarity of the situation.

Belief Bias. Another family of experiments demonstrating a common source of errors concerns *belief bias*, the human tendency to accept arguments with "believable" conclusions, and be reluctant to accept arguments with "unbelievable" conclusions, regardless of the arguments' logical (in)validity (Evans et al., 1983; Evans, 2003). For example, people tend to accept the conclusion of the following invalid argument, since it sounds believable given ordinary intuitions.

1. No addictive things are inexpensive.
2. Some cigarettes are inexpensive.
3. Therefore, some addictive things are not cigarettes.

That is, experimental subjects believe conclusions of arguments even though the underlying reasoning is problematic influenced by its prior believability, and do not believe other conclusions even when they logically follow from the premises. The results are even worse when subjects are put under time pressure, which has served as evidence for the competition between heuristics and slower, rule-based processes (Evans and Curtis-Holmes, 2005).

Suppression Task. The *Suppression Task* further demonstrates that people's performance does not adhere to the principles of classical logic (Byrne, 1989; Dieussaert et al., 2000). In particular, people suppress logically valid inferences (e.g. Modus Ponens and Modus Tollens) when additional premises are added. For example, subjects apply Modus Ponens with the following premises:

1. If she has an essay to write then she will study late in the library.
2. She has an essay to write.

...but fail to do so when the additional premise is added:

3 If the library stays open she will study late in the library.

The endorsement and suppression of invalid inferences (e.g. Affirmation of the Consequent, Denial of the Antecedent) also yield alarming results for the role of classical logic in reasoning.

Other examples. Difficulty with deductive reasoning is also exhibited in deductive reasoning games. For example, there is research on the deductive *Mastermind* game (Gierasimczuk et al., 2013; Zhao et al., 2018), focusing on which elements of the game are responsible for its cognitive difficulty. Moreover, there is evidence that decision-making is heavily influenced by the framing of options (Kahneman, 2011, Part 4), which can challenge properties discussed in Section 2.2 like *Closure under Logical Equivalence.* For instance, different responses are evoked whenever a question on the outcome of a surgery is posed in terms of survival or in terms of mortality, despite the equivalence of statements. This is called the *framing effect* (Tversky and Kahneman, 1985).

We now move to the schools of thought that have emerged from empirical evidence in order to explain the difficulty of real people in deductive reasoning tasks.

Mental Logic. Mental Logic is the view that people possess a system of formal, content-independent rules in their mind, which they employ in deductive reasoning problems. A well-known theory is Psycop, standing for *Psychology of Proof* (Rips, 1994). According to it, deductive reasoning amounts to finding mental proofs, i.e. generating sentences linking premises to conclusions. These links are provided by inference rules. The rules can be either *forward* or *backward* – depending on whether the subject proceeds from the premises to the conclusion or the other way round. This procedure does not always succeed. The number and the kind of the rules applied are indicative of a deductive problem's difficulty.

The asymmetry in endorsement rates for Modus Ponens and Modus Tollens, as well as the main response times of experimental subjects, can be explained using Psycop.[3] The system contains a rule for Modus Ponens (IF Elimination) but not for Modus Tollens. In other words, Modus Tollens is not a primitive rule, unlike Modus Ponens; deriving its conclusion requires an indirect proof, that also involves NOT Introduction and IF Elimination (hence the longer processing

[3] In (Marcus and Rips, 1979) the conclusion of MP is said to "follow" by 0.98 percentage (after 1907ms), as opposed to 0.02 "doesn't follow" (after 2119ms); the results for MT are, 0.52 (after 2882ms) and 0.48 (after 2245ms) respectively.

time). In this way, Psycop, much like the participants of the abstract selection task, derives that A should be turned (due to its IF Elimination rule) but runs into problem turning the 7 card (because of the additional requirements of Modus Tollens). The turn of card 4 might be explained by subjects assuming the converse of the conditional and following again IF Elimination. One explanation for the minority that succeeds is that these subjects considered explicit possibilities for the backsides of the cards, which triggered the application of the cognitively difficult and time-consuming Modus Tollens. Still, the content-sensitivity of the task and the facilitation effects of the variants are not directly accommodated by Psycop. Rips' defense identifies changes in memory as responsible for these effects and not changes in the mechanism of reasoning.

Mental Models. Contrary to Mental Logic, the Mental Models theory claims that people do not apply formal, content-independent rules (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991; Johnson-Laird et al., 1992). Instead, people construct models for sentences and read off tentative conclusions from these. This corresponds to two stages: *comprehension* and *description*. However, these tentative conclusions are inspected in another stage, called *validation*, that searches alternative models. Failures in reasoning tasks are then explained by assuming that the subjects read off a conclusion that is not true in all models of the premises. One of the claimed advantages is that this view is better equipped to deal with content effects, unlike Mental Logic.[4]

How are asymmetries between rules explained under this theory? Due to limitations in working memory, subjects follow an "economical" strategy, not representing all models explicitly from the beginning. Instead, they represent one model explicitly, and keep as much information as possible implicitly stored. For example, premise $p \rightarrow q$ has one explicit model where p cannot occur with not q, and an implicit one where p might not be the case. Then, if a second premise p is given, the first model is inspected, and the conclusion q is derived, which cannot be falsified. But if $\neg q$ is given as the second premise, then the explicit model will have to be eliminated and the implicit possibilities (all three mental models consistent with $p \rightarrow q$) will become explicit. Out of those, only the one representing $\neg p$, $\neg q$ is consistent with both premises, therefore the conclusion will be $\neg p$. Notice, that Modus Ponens, which is empirically considered an "easy" rule, required the check of only one model, while Modus Tollens, the empirically "difficult" rule, required three models. Johnson-Laird (2013) provides experimental evidence for the claim that the number of mental models is indeed correlated with the difficulty of a reasoning task.

Pragmatic Reasoning Schemas. Another view is that of Cheng and Holyoak (1985), which rejects the claim that the mechanism underpinning reasoning is

[4] But see (Rips, 1994, Chapter 10).

content-independent rules, but also the claim that the mechanism is based on representations like mental models. Instead, humans use *pragmatic reasoning schemas*, which are knowledge structures about a certain domain that contain generalized sets of rules. This is how the theory addresses the content effects of the selection tasks: we do better in those variants that fit a pattern abstracted from experience. Repeated experience induces principles that are content-dependent but the induction process that produces them seems to be content-independent. However, this proposal has been criticized for lacking a general, testable account of reasoning – focusing on specific problems like versions of the selection task but not explaining why people perform "easy" inferences without need for a pragmatic schema abstracted from experience (Rips, 1994, Chapter 9). Opposing theorists, such as proponents of Mental Logic or Mental Models, have further claimed that the results of this view are subsumed under their own theories.

Evolutionary Algorithms. Another approach argues that we have evolved a reasoning module in the course of our evolution that is suitable for problem-solving in *specific* domains: those that involve social exchange (Cosmides, 1989; Cosmides and Tooby, 1992). Even if it appears that we reason logically, we do so because the content of a task induced the use of these evolutionary algorithms. For example, the variants of selection tasks where subjects reason successfully are the ones that evoke algorithms suitable for "cheater detection". This approach shares with the previous one the view that people lack a mental logic and succeed in reasoning tasks whenever they apply inferences appropriate for the specific domain shaped by the problem. However, notice that this approach is entirely content-dependent, a feature criticized in (Rips, 1994, Chapter 9), and that social contracts appear to be only one of the kinds of schemas that could fit in the schema theory, a criticism given by Holyoak et al. (1995).

Experimental results and explanations vary on failures in deductive reasoning and they seem to hold a more or less favourable spot for logic in human reasoning. What is indubitably manifested though is that people are not perfect deductive reasoners, who immediately know/believe all consequences of their knowledge/beliefs. Moreover, their failures are *systematic* indicating that there is an underlying reason why deductive reasoning can sometimes be so difficult. Whichever is one's preferred analysis of what this underlying reason is, EL seems to be missing out on it. As we will see, this has implications for the normative (not only the descriptive) purposes of logical modelling as well.

2.3.2 Introspection

Apart from deductive reasoning, empirical research also challenges assumptions of unlimited positive and negative introspection. Studies on *implicit cognition* seem to corroborate the philosophical objections, showing that people can gain knowledge even without making a conscious effort to do so. For instance, subjects who learned from *examples*, managed to recognize well-formed strings of abstract languages, without thereby coming to know the underlying rule (Reber, 1967, 1993; Litman and Reber, 2005). Moreover, it has been documented that *implicit memory* affects our performance as we resort to our accumulated experience without making a conscious recall, unlike *explicit memory* that usually involves a deliberate and conscious act of recall (Schacter and Tulving, 1994). Naturally, the kinds of knowledge produced by such different processes would not obey the same introspection principles.

Similar evidence shows that people transfer skills from one complex problem to another analogically, without being aware of their shared attributes (Schunn and Dunbar, 1996). Experimental findings indicate that the access of subjects to complex mental processes, involved for example in judgment and decision-making, is not always infallible or direct (Nisbett and Wilson, 1977; Nisbett and Bellows, 1977). Moreover, subjects in the experiments of Johansson et al. (2006) seem to provide introspectively derived reasons on particular choices made in a decision task, while they fail to acknowledge mismatches between their intended choice and its outcome (*choice blindness*). Findings on implicit bias and evaluative priming are also taken as evidence for the fact that people have poor introspective access to their beliefs. For example, someone might implicitly hold a racist belief, while still not believing themselves to hold a racist belief (Lane et al., 2007; Schwitzgebel, 2010). Therefore, experimental results are in agreement with the objections already put forward by philosophers, as described in Section 2.2.2.

2.3.3 Reasoning about others

A crucial feature of real agents' higher-order reasoning is that of *Theory of Mind* (ToM), also known as *mindreading* (Apperly, 2010): the cognitive capacity to understand and predict external behaviour of others and oneself by attributing internal mental states, such as knowledge and beliefs (Premack and Woodruff, 1978). For example, when an agent knows or believes that the ball is in the box, the agent is using zero-order ToM. When an agent believes that Mary believes that the ball is in the box, she is using first-order ToM and so on. The case for attributing such mental states to one's self falls under the previous subsection. We now focus on attributing mental states to others.

Many experiments suggest that higher ToM levels are only achieved gradually in life and even adults' use of higher orders is greatly constrained – again, in stark contrast to the EL modelling, whereby agents perform attributions of knowledge/beliefs to others at any modal depth.

Let's delve deeper into the experimental evidence on limits of social cognition. The most paradigmatic experiments on ToM are the *False Belief Tasks*, asking subjects to discern their beliefs about reality from their beliefs about others' beliefs. A well-known task used to test ToM development in children is the so-called *Sally-Anne* task (Baron-Cohen et al., 1985). Successful performance requires discerning their own true belief from another's *false* belief. It turns out children on the autism spectrum and children younger than 4 tend to fail the task, reporting their own belief instead (Wimmer and Perner, 1983; Baron-Cohen et al., 1985). After mastering first-order ToM, children seem to develop their second-order ToM between 5 and 9 years old (Perner and Wimmer, 1985; Flobbe et al., 2008; Hollebrandse et al., 2014).

Similar limitations apply to adults. While the development of ToM continues, and can even reach the fourth order, adults too find it difficult to master tasks that require higher orders (Kinderman et al., 1998; Birch and Bloom, 2007). And while adults have developed the ability to tell apart their own mental states from others', they might still have problems systematically applying it. Verbrugge and Mol (2008) investigate the ability of subjects to use ToM in the context of the Mastermind game, finding limits on their second-order reasoning. Similarly, Keysar et al. (2003) show that adults have trouble reasoning about others in the context of a communication game. The participants had to move objects in a grid, and, unbeknownst to a person who acted as their "director", they hid an object in a bag. Then the director referred to one of the mutually visible objects with a term that resembled the hidden object. Most participants, despite knowing that the director does not know about the hidden object, moved the bag itself, and not the mutually observed item. According to Epley et al. (2004), people make judgments about others' perspectives by serially adjusting from their own, observing that adults were more biased by their own perceptive (*egocentric bias*) when put under time-pressure. The findings of Hedden and Zhang (2002) also exemplify the difficulty of adults applying ToM, this time in strategic matrix games: only some participants shifted to second-order ToM in the course of the game.

Marble Drop Games, logically equivalent to matrix games, have also been used to test the application of ToM by adults. For example, they have been used to investigate why adults, although expected to have mastered levels of ToM, fail in applying it when it comes to decision-making processes. In more detail, Meijering et al. (2010) show that the *context* of these games has a facilitative effect on the application of higher-order reasoning. Notice the analogy with deductive reasoning, i.e. the poor performance in the abstract selection task and the improvement in some of the variants. Meijering et al. (2011) further show that a supporting structure (in stepwise instruction, training, and asking for social reasoning) similarly improved performance. Therefore, the ability to apply ToM seems to be subject to improvement, rather than fixed once and for all depending on one's age.

However, additional cognitive costs are associated with the systematic application of ToM. This is supported by Meijering et al. (2013). The participants had to reason about another player; their performance turned out to be influenced by whether that other player was reasoning about the participant (*player condition*) or about a balance scale (*balance condition*). Both conditions required the same type of reasoning steps, yet the former required an additional perspective shift.

The participants' performance (e.g. their reaction times, which were shorter under the balance condition) suggests a difference in cognitive difficulty, which in turn indicates that perspective-taking comes with additional cognitive costs.

In the same spirit, Lin et al. (2010) emphasize the importance of cognitive resources in applying ToM: subjects with lower working memory capacity were less efficient in applying ToM than those with a higher capacity, and the ability to apply ToM was hindered when a secondary task was consuming the subjects' attention. Apperly et al. (2006) report on experiments whereby adult subjects responded more slowly to unexpected questions about another agent's beliefs about ontic facts than to questions about the ontic facts alone. The authors conclude that ascribing beliefs to others is not automatic, but it rather requires an effortful process that comes with its own costs.

The evidence for resource-consuming applications of ToM fits well in the landscape of studies on the reasoning strategies people follow, e.g. in the context of Marble Drop Games (Szymanik et al., 2013; Meijering et al., 2012, 2014). Given that resources are finite, it makes sense that players tend to follow an economical principle, prioritizing simple strategies for as long as they pay off and only resort to computing complex mental states when this becomes absolutely necessary.

2.3.4 Dual process theories of reasoning

The experimental results have led to diverse explanations on the underlying mechanisms of reasoning. Their repercussions, however, go even deeper as the interpretations of these results also touch upon the very understanding of "reasoning", which thus deserves special attention.

Stenning and van Lambalgen (2008) suggest that seeking *the* mechanism of reasoning might be a misleading goal. This is because it is important to specify the *level* (computational, algorithmic, implementational) in which the issue is to be studied (Marr, 1982). For example, the debate between proponents of Mental Logic and Mental Models might be empirically meaningless, because the argument patterns usually studied come from sound and complete logics, therefore manipulations with rules map to manipulations with models and vice-versa.

Aside from that, since experiments are intended to test people's reasoning, we need to specify what is exactly meant by it. The experimental results surveyed above have also given rise to the so-called *dual process theories of reasoning*. These theories suggest a richer picture wherein the empirical evidence should be interpreted (Stanovich and West, 2000; Evans, 2003; Kahneman, 2011).

According to these theories, there are different aspects in human reasoning, the result of two different mental processing types, often referred to as *System 1* and *System 2*.[5] Dual process theories characterize the operations of System 1 as

[5] Kahneman (2011, pp. 27-9) observes that Systems 1 and 2 are not systems in the standard usage of the term (sets of interconnected objects, or parts of the human brain). We stick to the

fast, automatic, intuitive, associative, unconscious, governed by habit, biases, and heuristics. The operations of System 2 have been characterized as slow, reflective, stepwise, rule-based, deliberately controlled, and effortful.[6] System 2 uses the inputs of System 1 to generate its outputs, following an orderly application of steps. When System 2 takes over, it engages in reasoning processes, of which deductive reasoning is one example, by breaking larger tasks into parts and generating our explicit knowledge and beliefs.

Given that the process is effortful, and our resources are bounded, it becomes clear that such processes eventually halt. This is in accordance with our experience (and documented evidence) of occasionally failing in demanding tasks due to cognitive overload. For example, dual process accounts suggest that the subjects of belief bias might try to reason logically and in accord with the instructions, but the influence of their prior beliefs (influence of System 1) is extremely difficult to suppress. So while System 2 has the ability to override and inhibit default responses generated by System 1, it requires high effort with respect to memory, attention, time, etc., which might prevent the override (Evans and Curtis-Holmes, 2005). Likewise, the selection task performance can be attributed to a System 1 generated bias, the matching bias mentioned above, especially manifesting itself under time-pressure (Evans and Lynch, 1973; Evans, 1998; Roberts and Newton, 2001). As we will see later, the dual process theories play an important role in assessing the significance of these findings within the Rationality Debate.

Although we have focused on specific kinds of processes (the ones EL is relevant for) and related psychological evidence, we have to underline that the two systems engage in a range of further activities. System 1, for instance, deals with face recognition, orientation, etc. System 2 deals with probabilistic estimates, the weighing of options, etc. In addition, while one of System 2's tasks is deductive reasoning, the dual process theories do not take a uniform stand on what its underlying mechanism is. In fact, as noted by Evans (2008), both Mental Logic and Mental Models seem to have mechanisms that map to the dual process account, accounting for both effortful deductive reasoning and for pragmatic influences.

2.4 The philosophical significance of empirical evidence

These empirical results clearly subvert the view of humans as "rational animals". But it is important to specify what exactly this view is and fix some terminological issues. There is a plurality of notions of "rationality" in the literature, often adapted to the parlance of the various scientific fields for which reasoning and

terminology, thinking of it as labelling families of processes.

[6] Also see the table in (Stanovich and West, 2000, p. 659) for a discussion of the systems as perceived by different theorists. A critical evaluation of the dual process theories can be found in Osman (2004).

decision-making are relevant.[7] In general, a behaviour is said to be "rational" whenever it matches a particular normative standard and is said to be "irrational" otherwise. In light of this, we can already identify two camps: the *Panglossians* and the *Meliorists* (Stanovich, 2012). The latter generally claim that reasoning is not as good as it can be (people's performance falls short of normative standards) therefore it can be improved, e.g. through education. The former argue for maximal human rationality; the descriptive matches the normative.

Yet, to set a solid ground for the Rationality Debate, and to understand its significance for philosophy and logic, one needs to specify which normative standard they have in mind, and thus against which standard they evaluate a behaviour as (ir)rational. The normative standards are usually implied to be those derived by (classical) logic and probability theory. This is the view assumed by different theorists (Braine et al., 1998) and it is identified as the traditional one by Cherniak (1986) and Stein (1996), upon observing that this view of rationality is usually thought to be the requirement for attributing beliefs and desires to agents.[8] The subject matter of the Debate is subsequently whether human reasoning competence (our underlying capacity to reason) matches the normative principles (how we *ought to reason*). In Stein's terms:

***The Standard Rationality Thesis* (SRT)**: people reason in accord with the rules of (classical) logic, probability theory, and so forth.

The normative principles within SRT are the ones derived from these rules, constituting the *Standard Rationality Picture* (SRP). In other words, SRT maintains that humans actually reason on the rails of SRP. Epistemic Logic adopts the norms of SRP for its agents; it provides a formal model for reasoners who develop their knowledge and beliefs in accord with the norms of SRP, i.e. a model for SRT-abiding reasoners. Contrary to that, an *Irrationality Thesis* maintains that people are irrational, i.e. that people do not reason in accord with these norms.

In light of the experimental evidence, even proponents of SRT have to admit that people do make mistakes. In what follows, we examine some of the defenses of SRT and we explain why we think these are inadequate (Section 2.4.1). Next, we consider whether an *Alternative Rationality Picture* should replace SRP as reference point, and what some difficulties in this project are (Section 2.4.2), before actually laying down what we think are features of a good alternative view (Section 2.4.3).

[7] The forthcoming *Handbook of Rationality* (edited by Markus Knauff and Wolfgang Spohn, the MIT Press) seeks to provide a comprehensive guide among this plurality of notions.

[8] This is also particularly important for the purposes of EL; for the independent interest of logicians, it is interesting to see what the implications of the psychological evidence are, when the normative principles are taken to be those determined by (classical) logic. This is because the reasoners formalized by EL are supposed to comply with these principles.

2.4.1 A critical review of arguments for maintaining SRT

Performance Errors. A popular reaction of proponents of SRT when confronted with the overwhelming empirical evidence is to treat mistakes as *performance errors*; failures to apply the correct reasoning step due to a momentary, random lapse.

To better understand this line of defense, we need a detour to linguistics. In language, like in reasoning, people who have the underlying knowledge of the subject matter (are able to read and write), still make occasional mistakes (Chomsky, 1965). These are hardly alarming for linguistic competence. All they indicate is some kind of interference (temporary memory failures, distraction, intoxication) that momentarily prevents proper linguistic behaviour. This point is crucial, for it discerns capacity to do something (e.g. ride a bike in Amsterdam or apply a logical rule) and failure to apply that capacity in only few occasions (e.g. after leaving from an (in)famous *coffee shop* in Amsterdam). Therefore, theorists argue that there is a fundamental difference between performance errors (mere mistakes, momentary lapses) and systematic divergence from the norms. The former do occur, but unlike the latter, do not threaten the idea of humans as rational beings.

Let's see this argument through an example. Cohen (1981) classifies failures in the Wason Selection Task as performance errors: the experimenters create an illusion by manipulating the situation in a way that prevents the subjects from reporting the correct solution. Listing the variants of the task where people perform better, Cohen concludes that the illusion created in the abstract task has to do with the question being so unfamiliar and artificial that subjects do not recognize what it really is about and thus fail in applying Modus Ponens and Modus Tollens, despite having the rules in their underlying competence. In this view, subjects do not lack the correct principles, but they rather make a performance error under certain circumstances.

However, invoking the distinction does not suffice to maintain SRT. In lack of an independent argument on why performance errors are all there is in failure, this argument amounts to an *ad hoc* immunization strategy (Stein, 1996): whatever people do matches the normative principles and if we get experimental results to the contrary, then they must be due to performance errors.[9]

More specifically, Stein (1996) emphasizes that the transfer of the argument from linguistics lies on shaky ground: performance errors in language have plausible explanations and a concrete pattern they fit in. The competence/performance distinction is not invoked just whenever empirical data do not fit in the prevalent theory. What we would need to apply the analogy is an independent reason to justify that, for example in the Wason Selection Task, Modus Ponens and Modus Tollens truly are in the agent's competence. What Cohen seems to claim is that

[9] As Kahneman (1981, p. 340) observes, there seems to be a kit of defenses whenever reasoning errors are observed (such as temporary insanity or a difficult childhood), that will allow theorists to restore the presumption of rationality.

the fact that subjects sometimes succeed in variants of the task qualifies as an "independent reason". But it can well be that the principle subjects use is not the one SRT predicts (a logical rule), but rather one that happens to give correct results solely for these and only variants (e.g. a pragmatic reasoning schema or a cheater detection algorithm), or because there are the right situational circumstances in the cases of success. Moreover, the distinction alone is vulnerable to the fact that subjects fail *systematically*. This is evinced by the empirical findings. If mistakes really are *random*, momentary lapses due to carelessness, intoxication, temporary distractions, then there should not be any correlations among them across tasks. Against this, Rips and Conrad (1983) report on experimental results on individual differences in deductive reasoning which reveal that subjects' scores on propositional tests are correlated with their performance on reasoning tasks. Stanovich and West (2000) list tasks – including selection tasks and belief bias problems – where such cross-task correlations and internal consistency of failures within tasks can be observed.[10] In conclusion, independently motivated arguments to back the performance errors account are necessary, if it is to lend support to SRT.

Computational limitations. Another reason why people, despite being rational, might not act in accord with the norms is their inherent computational limitations. Reasoning does not occur in a vacuum, but in a context shaped by our cognitive resources (like memory, attention) and situational constraints (e.g. time pressure). It is unfair to say people are irrational because they sometimes fail in reasoning tasks, or as Stich (1990) puts it, "it seems simply perverse to judge that subjects are doing a bad job of reasoning because they are not using a strategy that requires a brain the size of a blimp" (p.27). Indeed, cognitive capacity with respect to certain resources, like working memory, is correlated with deductive reasoning performance (Bara et al., 1995).

Facts on computational limitations are acknowledged by all camps in the Rationality Debate. Some theorists utilize these observations to argue that people have evolved to reason correctly within their limitations (Gigerenzer, 1991). In their view, reasoning *is* as good as it gets, therefore the participants of the experiments should not be accused of irrationality and the empirical evidence does not have any devastating implications for the view of people as rational.[11]

A full-fledged proposal on the basis of limitations is presented by Cherniak (1986), who appeals to the *Finitary Predicament* of humans: we have a fixed limit on our neurons, lifespan, cognitive capacities, time available for reasoning. As a

[10] The authors also observe that an emphasis on systematicity is also used (e.g. by Thaler (2012)) against proponents of perfect market rationality, who often argue that the mistakes people make do not affect the bigger picture, for they are random or tend to cancel out. Evans (1984) also adopts the distinction between random and systematic errors to clarify the notion of a bias, that is capable of threatening SRT.

[11] In other words, the discrepancy between the descriptive and the normative is insufficient for a charge of irrationality, because it does not qualify as a discrepancy between the descriptive and the *prescriptive*, i.e. what people should do given their limitations.

result, ideal rationality should be replaced by a standard of *Minimal Rationality*, wherein an agent undertakes *some*, but not necessarily all, of those actions which are apparently appropriate. This in turn translates to her ability to make inferences and eliminate inconsistencies. In this view, the subjects of the experiments are not irrational, sloppy, and careless. In a cost-accuracy trade-off, experimental subjects might actually do a relatively good job.[12] For illustrative purposes, consider the following principle, which sounds reasonable for a SRT-theorist, adapted from (Cherniak, 1986, p.17):

> If you have a particular belief set, then if any inconsistency arose in the belief set, you would eliminate it.

But considering the Finitary Predicament, it is clearly impossible to follow this procedure every time we revise our beliefs. It would actually be practically useless to waste a lifetime's resources inspecting our belief set for consistency. But if, in light of this, we resort to Minimal Rationality as a superior standard, the crucial question is how those "*some* actions/inferences/inconsistencies" can be determined. Cherniak argues for a theory of feasible inferences and a theory of human memory structure, designed with the help of experimental psychologists.

At face value, arguments that stress the importance of computational limitations as explanations of the performance of experimental subjects do alleviate the accusations of irrationality and the pessimism on the potential of human reasoning. However, they do so while not supporting the standard picture.[13] That is, they – directly or indirectly – suggest that the standard picture should be replaced by another picture that is better aligned with limitations of finite human beings.

Interpretation. Another way to defend SRT is to attribute mistakes to misinterpretation. The subjects might be interpreting the task differently, compared to the experimenter's intentions, and thus not give the correct response. Therefore, the experimental results do not test what they are supposed to be testing (the reasoning competence) and the subjects only *appear* to be reasoning incorrectly. In Henle's words: "where error occurs, it need not involve faulty reasoning, but may be a function of the individual's understanding of the task or materials presented to him" (Henle, 1962, p.373).

These mistakes are sometimes subsumed into the aforementioned accounts (Cohen, 1981). But, given that performance errors are random, misinterpretation

[12] Cherniak goes further, arguing that it would be irrational to even try to satisfy ideal rationality conditions. This is reminiscent of Evans et al. (1993), where different notions of rationality are distinguished: logicality and goal-oriented rationality. Conforming to the former does not necessarily imply conforming to the latter.

[13] Unless, as Cohen (1981) says, we take deviations due to computational limitations as mere performance errors that say nothing about underlying reasoning competence. Throughout this subsection though we argue that this type of deviations should not be taken as performance errors.

mistakes would not qualify as such as they are expected to occur as long as the source of misinterpretation persists. Moreover, subjects who fail because they run out of the resources might still have the correct (intended) interpretation of a task.[14] This view acknowledges that the responsibility behind imperfect performance may be shared equally between both subjects and experimenters.

Indeed, experimental findings show how deviations from normative standards could be seen as appropriate responses, but to alternative task construals (Tversky and Kahneman, 1983; Gigerenzer, 1991; Hertwig and Gigerenzer, 1999). In this spirit, Stanovich and West (2000) observe that when alternative construals are off the table, cognitive ability differences are mitigated. The findings behind the dual process theories discussed in Section 2.3.4 are crucial in identifying the source of (mis)interpretation, as the two systems give rise to different interpretations. This distinction may explain the difference in performance among variants of the selection task, e.g. versions that rely on the automatic workings of System 1 (like the deontic ones) or on the effortful computations of System 2 (like the abstract ones).[15]

Taking interpretation and the role of the two systems seriously, but without committing to SRT, is advocated by Stenning and van Lambalgen (2008).[16] There are two stages in reasoning: *to* an interpretation and *from* an interpretation. Both involve reasoning, and neither should be neglected. The latter usually embodies what is supposed to happen in a reasoning task: taking some explicit premises, the subjects reason deductively (a System 2 task) towards a conclusion. But since premises are given in natural language, an interpretational process occurs (namely, discourse interpretation, a System 1 task) and it is sensitive to certain parameters. This parameter-fixing might yield a different, not necessarily classical, logic. Dismissing logic's place in reasoning altogether may simply be the outcome of understanding logic as necessarily classical logic and disregarding the parameters involved in the interpretational stage. This point is relevant for both empirical scientists and formal modellers, like epistemic logicians: the former should abandon a narrow understanding of "logic" and the latter should not neglect either reasoning stage in modelling attempts.[17]

Overall, the misinterpretation account seems to hint at some modification of the standard picture as well, instead of lending direct support to SRT. Unless one insists on taking misinterpretation mistakes as performance errors, in which case,

[14] The literature generally does not assume an absolute distinction between the arguments in the way presented here. There is often an overlap in the way theorists appeal to performance errors and misinterpretation. Apart from the reason mentioned above, we choose to treat the arguments separately because their subtleties dictate different stances to SRT.

[15] Still, System 1 can sometimes be sensitive to logical cues (Bago and Neys, 2017).

[16] See (Counihan et al., 2008) for interviews of selection task participants, backing this claim.

[17] In this way, the view of Stenning and van Lambalgen (2008) reserves a milder spot for Piaget (1953)'s formal operational stage (which adults are supposed to have reached, thus mastering classical logical inferences) against the criticisms ensuing from Wason's findings.

the account inherits objections targeted at proponents of the performance errors.

We have argued that the attempts to defend SRT by explaining away the empirical evidence as performance errors, computational limitations, or misinterpretation are not sufficient. The first because it needs further independent arguments to not be a mere immunization strategy; the second and the third because, despite offering very useful insights in the analysis of experimental results and defending subjects against extreme charges of irrationality, they defend RT not in terms of the standard picture but by implicitly suggesting another. But before we depart from the standard picture, are there any independent arguments for the performance errors account to restore its credibility and strike back? In what follows, we review some of them.

Charity. In order to treat experimental failures as mere performance errors, one can invoke the *principle of charity* (also known as the principle of *rational accommodation*). Quine introduced it as a guide for language translation (Quine, 1960). The rough idea is that if your interlocutor speaks in another language and says something seemingly silly, you would probably not question her rationality, but rather your own translation. This idea behind translating utterances can be adapted to interpreting principles used in reasoning (Davidson, 1973; Dennett, 1987, 1981). If subjects of a reasoning task perform in a seemingly deviant way, we should not question their rationality, but rather our interpretation of it. Proper interpretation of someone's reasoning requires the rationality of that person. Dropping SRT then amounts to never attributing beliefs to people!

But the view that intentional description of others' mental states presupposes perfect (SRP-wise) rationality is under fire due to the Finitary Predicament (Cherniak, 1986; Stich, 1990). Charity is not compatible with the inevitable limitations of human cognition. Our reasoning sometimes simply does not live up to the normative principles, no matter how charitable the interpretation is. Charity alone does not offer any explanation for that.

Thagard and Nisbett (1983) ask for charity in moderation, leaving room for empirically indicated judgments of irrationality. Yet Stein (1996) attacks *Weak Charity*, the view that people should be interpreted as rational, unless there is strong empirical evidence to the contrary. This cannot support SRT because the real challenge posed by the empirical evidence is exactly that there *is* strong evidence to the contrary. Another principle that is sometimes adopted, instead of Charity, is *Humanity*, which does not fine-tune translation to the rationality of the translatee (like Charity), but fine-tunes it to the extent it establishes agreement between the translatee (the subject) and the translator (the experimenter) (Grandy, 1973). But it is hard to see how this helps: as Stein (1996) observes, it might well be that it is the experimenter who is irrational, so there is no reason to see agreement as support for SRT. Another alternative is Holistic Charity: we should interpret people's behaviour *holistically* and be considerate of the context in which it arises (Davidson, 1967). This leaves room for subjects who might fail due to deficiencies in their underlying competence, as long as these failures are interpreted in the context of an account that sees humans as *mostly* rational. For example, people might employ a heuristic (System 1) that is mostly, but not entirely, effective when resolving a reasoning task. However, it is not clear what constitutes much or little (ir)rationality, hence this point is too vague to safeguard SRT by itself.

Another possible fix is through combining Charity and Minimal Rationality. In this view, to attribute beliefs to someone, is not to say that they are rational, but *minimally rational* (in Cherniak's sense). For example, people might make some inferences but not all of them, e.g. Modus Tollens in the Wason task. While this response is more moderate and better aligned with experimental evidence, it fails to lend support to SRT, and at best it suggests we revise the S in SRT.

A way to counter this objection might be to sweep limitations under the rug of performance errors, which are admitted without harm to SRT. We have already argued against this tactic. But more importantly, notice the question-begging flavour of defending Charity using performance errors, when Charity is needed as an independent argument for why mistakes can only be mere performance errors.

Reflective Equilibrium. Another conceptual argument for SRT is based on *Reflective Equilibrium* (RE) as a theory of justification (Daniels, 2020). Applied to reasoning, the argument goes: since both the normative principles of reasoning and the principles in our reasoning competence are formed by our intuitions on what constitutes good reasoning, they cannot diverge. In the words of Goodman:

> This looks flagrantly circular. I have said that deductive inferences are justified by their conformity to valid general rules, and that general rules are justified by their conformity to valid inferences. But this circle is a virtuous one. The point is that rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either. (Goodman, 1955, p.64)

That is, the normative and the descriptive converge because they originate from the same process (reflective equilibrium) using the same inputs (intuitions). So any observed divergences must be mere performance errors. In Cohen's words:

> [W]here you accept that a normative theory has to be based ultimately on the data of human intuition, you are committed to the acceptance of human rationality as a matter of fact in that area, in the sense that

> it must be correct to ascribe to normal humans beings a cognitive competence – however often faulted in performance – that corresponds point by point with the normative theory. (Cohen, 1981, p.321)

Still, the view that the normative principles of reasoning come from a process of RE with our intuitions as input is vulnerable to empirical evidence. Subjects systematically violate the principles of classical logic and probability theory so principles that are in RE for people, might well be not so rational at the end. For example, a principle that seems to be in RE for many people is the *gambler's fallacy*, but nonetheless it is considered irrational given our normative standards.[18]

There are however certain modifications of RE that are meant to address this crucial objection. The rough idea is to impose a filter on which intuitions really count. First, one might say that it is only the *considered*, and not the *naive*, intuitions that play a role in the rise of these normative principles.[19] But given that mistakes are systematic and that the subjects often insist, even after repetitions of the tasks or presentation of contrary evidence, there is no reason to assume this modification will block all objections. Most importantly, it is not clear what draws the line between *considered* and *naive* intuitions, and whether this can be defined in a non-question-begging way. A second way to filter intuitions is to appeal to the epistemic authority: the intuitions that truly matter are those of a specific subset of people, the *experts* (Stich and Nisbett, 1980).[20] Still, this filtering only transfers the question-begging flavour to the definition of the "expert". Moreover, experts are themselves fallible human beings. As Stich says, we cannot eliminate the possibility that an expert, who is still a fallible human influenced by ideology or recreational chemistry, ends up endorsing a nutty set of rules (Stich, 1990, p.86). This is in fact empirically vindicated since the experimental subjects are often people with good formal education and logical training.

Another way to prevent non-rational principles from sneaking in as normative principles is by telling apart *narrow* and *wide* RE (Cohen, 1981; Rawls, 2009; Daniels, 2020). Wide RE asks that we also need to bring our general background philosophical theories into agreement with our intuitions. However, considering gambler's fallacy, Stich (1990) argues that there is no reason why a gambler who accepts the fallacy will cease to do so once facing the philosophical considerations against it, or that general philosophical theories, however defined, will not be themselves guilty of importing faulty principles. Thagard (1982) observes that when background psychological theories are also included in the process, the

[18] The fallacy describes the belief that if an event has occurred more(/less) frequently than normal, it is less(/more) likely to happen in the future. One can even find 19th century logic manuscripts in which versions of the fallacy are endorsed (Coppee, 1860), as Stich and Nisbett (1980) observe.

[19] Notice though that Cohen's view stresses the importance of *naive* intuitions when it comes to reasoning (Cohen, 1981, 1994).

[20] Thagard (1982) argues on these grounds that Cohen (1981), who focuses on naive intuitions, follows a *populist strategy*, as opposed to the *elitist* strategy of Stich and Nisbett (1980).

empirical evidence can also play an important role especially concerning people's physical and psychological limitations.

Another provocative criticism against RE, developed by Stich (1990), is that this process is a sign of *epistemic chauvinism*, that lacks intrinsic or instrumental value. Stich sees no reason why the evaluative notions embraced in our culture are superior to the ones of some exotic folk. There is no special reason why grounding this process on our intuitions and philosophical theories would alleviate our concerns, especially in light of experimental results manifesting a variety of inferential practices (e.g. inconsistency-tolerant (Priest, 2006)).

Concerns may also be raised for the assumption that the descriptive part, i.e. the principles we do have in our competence, are generated from a process of reflective equilibrium (Stein, 1996). This presupposes that introspection is an infallible and unique source of data. However, introspection is itself empirically challenged as an (infallible) source as we saw before. It is not unique either: there are other inputs in our process of investigating human reasoning competence. An insistence on RE disregards the persistent interdisciplinary findings – psychological, evolutionary, computational – that are relevant in this investigation. Yet admitting them into the picture again goes against the very gist of the *standard* rationality picture. We would have to either smuggle empirical considerations as playing a role in the descriptive part (and thus acknowledge that there might be a gap between the descriptive and the normative – RE goes astray) or smuggle empirical considerations into the normative part as well (and thus distort the SRP – SRT goes astray). Once again, in looking for ways to defend SRT, we are confronted, at best, with the need to revise it.

Evolutionary arguments. Another family of arguments consists of evolutionary arguments that appeal to natural selection to argue for the existence of limits on human irrationality (Quine, 1969; Dennett, 1981; Fodor, 1981; Sober, 1981). They are roughly based on two premises: (i) evolution produces close approximations to optimally well-designed systems, and (ii) optimally well-designed cognitive systems are rational (Stich, 1990). Stein (1996), however, disagrees that the filter of natural selection is fine-grained enough to aim at logical truths and consistency preservation, as the standard norms ask. Stich attacks evolutionary arguments, by highlighting misunderstandings on evolution and natural selection that give rise to this pattern of argumentation. Moreover, even if we treat natural selection as a faultless optimizer, the application of the argument to reasoning neglects social influences and culturally inherited practices. Such elements lend support to a view of cognitive pluralism, that challenges SRT, even if grant that natural selection provides an innate basis for cognitive processes that we all share.

2.4.2 Changing the norm?

In Section 2.4.1, we showed that SRT cannot survive the challenge posed by the empirical evidence; attempts to defend it either reach a dead-end or, at best, suggest it should be revised. In light of this, one might simply claim that we should replace the standard picture for it is inadequate or obsolete. A strand of Panglossians attempt to restore the match of the normative and the descriptive by rejecting the existing norms and advocating for an alternative reference point, which would be better suited as a criterion for human rationality. But why should we engage in the quest for an alternative picture and not just accept that people are irrational since they systematically fail in some reasoning tasks? Are there any arguments for RT that simultaneously reject (or not commit to) SRT? In answering this question, we also locate some challenges of the quest for a change in norms.

Stein (1996) criticizes a *no-access* argument (Macnamara, 1990), according to which even if we were irrational, we would not be in a position to know it. This is because we *cannot* possibly have any access to what the normative standards are, because in investigating this we use the same capacities that we seek to assess. Therefore, we cannot credibly state that normative standards diverge from the experimental observations, and thus, that people are irrational. In other words, the Irrationality Thesis is epistemologically inaccessible. Hence some form of the RT must hold, but for the same reason, we cannot specify it. Still, this is a negative argument against Irrationality and it does not provide adequate support for an Alternative Rationality Thesis (ART). This is because Irrationality might be inaccessible, but not necessarily false. Moreover, it does not provide a *constructive* resolution of the Rationality Debate.

Another view is heavily influenced by the computational limitations of human agents. According to it, the norms against which we assess the performance of real people should be indexed to human reasoning ability. This means that, as far the experimental subjects reason as well as they can, they are rational; asking them to comply with the standard picture is simply making an unrealistic demand of them. In this spirit, Levi (1983) and Messer and Griggs (1993) claim that it is the researchers who commit the fallacy (by using a wrong normative model to assess performance), and not the experimental subjects. This is in fact reminiscent of the "*ought implies can*" principle in Ethics: if you have a moral obligation to perform an action, then you must be able to perform it. But if we are to replace the standard picture, what should be the replacement? This new picture should be one that prescribes behaviour actually attainable by human beings, that is, a more "pragmatic" picture of rationality.

There are certain challenges in this project. First, the boundaries of pragmatism are vague. Just asking for a "pragmatic" picture, without specifying concrete criteria, does not qualify as a constructive alternative. Second, this project only renders people safe from a charge of irrationality, provided we show that experimental subjects are indeed reasoning as best as they can. This, dubbed the *optim-*

istic picture of the human competence by Stein (1996), calls for more justification. While for some failures of the standard picture, this might be straightforward, for others it is not. Why is it that subjects who fail to apply Modus Tollens do as well as they could possibly do? Why is it that mathematicians do as well as they could possible do, despite Goldbach's Conjecture remaining an open problem?

One way to deal with these problems is to claim that the norms are the ones that are directly indicated by the empirical evidence. For example, Gigerenzer (1991) suggests a return to the "good old days" when modellers, upon observing deviations between their models and human performance (e.g. due to St. Peterburg's paradox), would be more likely to revise their own models than blame the experimental subjects. In this view, which resembles the RE argument, people are necessarily rational.[21] This still preserves the question-begging flavour of RE. We can imagine scenarios where there is another principle that should be endorsed and subjects simply fall short of it (just recall the gambler's fallacy). This argument eliminates the possibility of evaluation. The matching is not established because we have independent, constructive reasons to define norms as such-and-such and we observe that they indeed match the actual performance, but rather because in order to necessarily restore rationality, the definition of the normative principles collapses to the experimental performance of the majority. Finally, this does not provide a good explanation for why the standard picture and performance sometimes *do* coincide; then, there seems to be no objection to the methodology and predictions of the modellers.

Another view criticizes SRT for ignoring the cognitive diversity of real people and suggests that normative principles be tailored to *each* human's capabilities.[22] This could be seen as a version of pragmatic rationality. But it still comes at the same shortcomings, e.g. not allowing for the possibility of evaluation, as it too derives *ought* from *is*. What we need instead is a principled way to provide indexed normative standards in a way that does not beg the question.

2.4.3 In search of an alternative picture

But then what can be an adequate picture to replace the standard picture, while still overcoming the challenges identified above? To answer this, first notice there are features of the standard picture that can be thought of as desirable. For example, it is precise, i.e. it indicates exactly what the normative principles are. These principles cohere well with our knowledge in disciplines like mathematics

[21] Notice though, that while RE fixes the norms and argues that competence cannot possibly fall short of them, what this ART argument says is that we should keep the reasoning competence fixed (at least what the experimental majority does) and tailor the normative standard around it. While both Cohen (1981) and Gigerenzer (1991) argue for the match between the two, their views have different starting lines.

[22] This is also a line advocated by Stich (1990), who discerns (descriptive and normative) notions of cognitive *monism* and *pluralism*.

and logic, and they allow for the possibility of evaluation, i.e. they do not read off *ought* from *is*. Stein (1996) observes that it is a desirable feature for our theories across disciplines to cohere well with one another. From the logician's perspective, it is especially interesting to see the established progress in logic cohere well with its use in the modelling of human reasoning and the Rationality Debate.[23] In short, our alternative normative principles should preserve such features while respecting the inherent limits of human cognition and surviving accusations of epistemic chauvinism.

Stein (1996) argues for a *naturalized picture of rationality*, that includes scientific evidence, next to philosophical theories, general intuitions, and first-order judgments on what constitutes good reasoning, as inputs in a *wide* RE process. Taking scientific evidence as input for the wide RE process, addresses the objection of Finitary Predicament. For example, scientific evidence might filter out a behaviour that would be suggested by simple RE processes (e.g. the ones that yield the standard picture), but is provably unattainable due to cognitive limitations. It also addresses objections of epistemic chauvinism because background philosophical theories and scientific evidence can counter-balance potentially chauvinistic inferential practices. No unique group's intuitions determine the alternative picture. However, this does not preclude evaluation, as the norms are not read off from actual reasoning performance. This is not the only input in the RE process. For example, it can be that experimental evidence indicates a prevalent behaviour, e.g. the gambler's fallacy, that is nonetheless not endorsed as a norm due to philosophical considerations.[24] This naturalized approach to rationality results in a naturalized approach to epistemology:

> [A] naturalized epistemology is an approach to the theory of knowledge that tries to develop an account of how we ought to arrive at our beliefs, while allowing, in contrast to traditional epistemology, that empirical facts can play an important role in this inquiry. The naturalized picture of rationality is a part of naturalized epistemology because it says that empirical facts play an important role in trying determine which principles of reasoning we ought to follow. (Stein, 1996, p.265)

Similar remarks are made by Goldman (1978), advocating for a shift from traditional epistemology to *epistemics*, which operates in a close alliance with psychology of cognition to provide advice about intellectual operations. This project unfolds in three directions. First, avoiding overly simplistic models that deal

[23] For a discussion on logic, epistemology, and the unity of science, see Rahman et al. (2004).

[24] Notice that proposing an alternative picture of rationality does not entail endorsing the respective rationality thesis or irrationality thesis. This remains an open issue given the ongoing research studying human reasoning competence. The more this progresses, the better insights we have to resolve the question. The interest in alternative normative standards is crucial for the purposes of epistemic logicians, who seek formalizations corresponding to rationality norms.

with the "beliefs" the agents should have; this mental classification is too coarse-grained to do justice to our rich cognitive life. Second, taking agents' cognitive capacities seriously in order to design normative principles attainable by real cognizers. Third, identifying the sources of flaws or defects of our cognitive system.

While in agreement with Stein's proposal and Goldman's motivation, we should be cautious on what is meant by *reasoning* when addressing human reasoning competence, even when the latter is evaluated against an alternative picture. In Section 2.3.4 we emphasized that there are different types of mental processes, which naturally come with their own individuated constraints. Stein's talk of an alternative picture hints at an all-encompassing reasoning system but its details seem to reflect exclusively System 2 processes (or in the terms of Evans et al. (1993), exclusively rationality$_2$ (logicality)). This neglects the influence of System 1 processes. Granted, the findings behind the dual process theories could be thought of as scientific evidence imported in the wide RE. Still, evidence for the existence of automatic and implicit processes threatens the status of the alternative picture as a normative standard: to say that an agent ought to reason in one way presupposes that she exercises deliberate control of the process. But this cannot be said for this type of processes. Therefore, the alternative picture should do justice to the nuances of different systems, taking into account evidence on the Finitary Predicament – which focuses on our explicit reasoning – but also accepting that its inputs are often generated by a different family of processes. Both features, i.e. our cognitive limitations and the variety of underlying processes, are central to the alternative picture we endorse.

This might sound similar to the project of Stich (1990) for *epistemic pragmatism*, which draws inspiration from Cherniak's ideas on feasible inferences and the structure of human memory, and utilizes findings from other fields, from cognitive science to complexity theory. Pragmatism, unlike chauvinism, takes on board interdisciplinary scientific findings and shapes an empirically plausible picture over which values should guide human reasoning. Yet the pragmatist project does not hold a favourable spot for the notion of *truth*. Truth does not have any intrinsic or instrumental value, i.e. it is not included in the diverse goals one's reasoning aims to achieve. Notice that naturalizing rationality does not entail this (harsh) stance to truth: truth need not be filtered out from the wide RE process Stein discusses.

Looking more critically into this, Solomon (1994) observes that the experimental deviations from standard norms do not really warrant the claim the people do not value truth and are instead better at realizing pragmatic goals other than seeking the truth. On the contrary, we often make mistakes in realizing very pragmatic goals, such as estimating future utilities or taking health-related decisions (the handling of the COVID19-crisis is perhaps a case-study attesting to that). Besides, if we were to abandon the quest for truth, in a reading of pragmatism as mere wishful thinking, the consequences would be rather unpleasant for the very purposes pragmatism supposedly serves. In a similar spirit, Harman (1991) circumvents the arguments of Stich (1990) by rebranding the claim that one cares of beliefs being true, which Stich attacks, to a claim that one cares that one's desires come true, which is more appealing to a pragmatist.

Overall, while pragmatism endorses the role of scientific evidence, our alternative picture, closer to Stein's understanding of RE, acknowledges the epistemic virtue of maintaining true beliefs. Consider for example *consistency preservation*, which seems desirable from the perspective of intuitions and background philosophical theories. Yet always maintaining a consistent set of beliefs is practically impossible, as demonstrated by scientific evidence, from cognitive science to computational complexity.[25] Because of the former point, it makes sense to ask that no explicit contradictions, of the form p and $\neg p$ are believed. However there can be implicit contradictions in our belief set, such as $\{p \to q, \neg q, p \vee r, \neg r \vee q\}$, which require effortful reasoning steps to be uncovered as such. To properly fix the extent to which consistency preservation qualifies as a norm, one should take into account the scientific evidence both (a) on the reasoning steps and the effort they require with respect to our time, memory, attention, etc., and (b) on what might have given rise to inconsistencies, e.g. one might entertain them due to the effect of belief bias.

2.5 Epistemic Logic and its place in the debate

2.5.1 Revisiting the received view

Epistemic Logic, as presented in Section 2.2, adopts the norms of SRP and provides a formalization that corresponds to SRT-reasoners. We have critically reviewed arguments for SRT and concluded that it is necessary to revise SRP and revisit the Rationality Debate. Therefore, the received logical view also inherits these criticisms. More specifically, the received view is clearly in conflict with the aforementioned findings: it predicts that agents are perfect deductive reasoners, contrary to the findings of Section 2.3.1; it predicts that agents introspect unlimitedly, contrary to the findings of Section 2.3.2; and it finally predicts that agents can make attributions to other agents at any modal depth, contrary to the findings of Section 2.3.3. Despite these discrepancies, there have been attempts to defend the standard EL modelling, independently from the SRT-criticisms. Below, we reflect on these lines of defense.

Line of defense 1: Idealization as in natural sciences. Motivated by examples from natural sciences, e.g. the use of frictionless planes in physics, the idealizations of the standard modelling can be defended as the means to reach the mechanisms underpinning the complex theory of knowledge and belief (Stalnaker,

[25] As Cherniak shows, having only 138 beliefs would make it impossible to maintain consistency even for a supermachine. In other words, this is an intractable task. A full-fledged explication of what has been called the "*tractable cognition thesis*" is offered by Van Rooij (2008).

1991; Godfrey-Smith, 2009; Frigg, 2010). Moreover, ignoring some features may be justifiable because the internal dynamics tend to move the system in question towards an equilibrium. Idealization in models of human reasoning can be justified by viewing the fallibility of agents as a kind of "cognitive friction" that interferes with reasoning processes yet the latter tend to reach an equilibrium of perfect rationality. Even if the equilibrium is never actually reached, the study of this ideal condition helps us understand these processes better. Take consistency preservation: while agents might not be able to reach perfect consistency in their set of beliefs due to "cognitive friction", they at least try to approximate this ideal condition. Idealization serves an evaluative purpose in which, however humans fail, their ultimate goal is to approximate the standard predicted by the mainstream proposals: the closer, the better. Weisberg (2007) draws distinctions between kinds of idealization and explains how they are grounded on the type of activity that gives rise to them, the justifications and the trade-offs involved in modelling. Yap (2014) also extends considerations from idealization in science to idealization in logic.

However, resorting to the idealized models of other disciplines is a convenient, but not fully accurate, analogy. It gives no clear reason, theoretical or empirical, to assume that the reasoning process, constantly influenced by external information and internal limitations, can ever reach an equilibrium of spotless rationality. Moreover, the indeterminacy involved in *what* counts as a good approximation weakens the effectiveness of this type of modelling. The internal coherence of experimental subjects and the systematicity in their errors shows that there is more to their reasoning processes than trying to reach an equilibrium, *modulo* some random slips. There is an additional difference: once natural scientists manage to account for more realistic assumptions, their new models are more reliable and accurate and they are therefore preferable to the earlier, more idealized ones. Contrary to this, the defense of the standard logical systems seems to presuppose that they fulfil their purpose at their *current* state. For example, it is the full form of consistency preservation that is considered the ultimate EL standard of evaluation, rather than a moderate version of it, tailored to empirical evidence. As a result, one cannot do away with the discrepancies by claiming that analogies to natural sciences prove there is no threat in the long run.

Line of defense 2: Simplicity. A second reason backing idealization lies in the need for simplification: there might be distortion, but its cost is unimportant when compared to the benefits of simplifying (Stalnaker, 1991). For example, ignoring some complicated elements of reasoning processes could be justified if (a) it contributes to a modelling that still reveals structural properties of the system or helps derive important conclusions, and (b) the incurred distortion is negligible. Besides, idealizing for the sake of simplicity has already contributed to a better understanding of otherwise vague philosophical notions. Admitting no simplification to save logic from any charge of idealization would only result in a logic that gives no interesting, illuminating, or distinctive results. Simplifying serves as the starting point: the modellers need to start *somewhere*.[26] Moreover, there are pragmatic benefits in this, as explained by Yap (2014): idealizing assumptions can have a constructive role, e.g. when made for the sake of tractability, and they can still give insight to actual phenomena. Besides, EL is not a static field and objections put forward against it are not to disregard it altogether but to delineate avenues for further refinement. In this sense, the question is *which* idealizations are good, and not whether they should or should not be used at all.

However, the distortion is not always negligible. The importance of the psychological evidence lies exactly in that it overshadows many benefits of simplification. For example, ignoring evidence that is so essential in the Rationality Debate is not a harmless assumption, for it distorts what logicians mean when they claim to have a logic for knowledge or belief of human agents. Yet formal modellers are right in emphasizing that the crucial point about simplification is not *whether* it should be used, but *how* it should be used. As a result, we need a more nuanced view when generally assessing the appeal to simplification. We view its role as counter-constructive, when it is invoked as panacea to silence any mild concern in light of the scientific evidence and to establish the traditional view as the ultimate one, resorting to the solitude of anti-psychologism. On the other hand, if simplification is seen as the starting point of modelling, providing a promising basis for further developments, suggesting avenues for future research, clarifying concepts and connections between them, then its role is important and scientific evidence is not a threat but an ally in this enterprise. Overall, simplification has indeed a very fruitful role in logic (much like in the natural sciences), but it can only be constructive when it is free of dogmatic forces that make it oblivious to what other disciplines have to say on human reasoning.

Line of defense 3: Normativity. Another source of justification is presented in terms of normativity. Although the standard logics draw an ideal picture, they are valuable as they set the standard that rational agents *ought to* comply with. Colyvan (2013) spells out this view when considering the justification behind these models, and others, e.g. in decision theory. The author agrees that models, when seen as descriptive, are not accurate as there is overwhelming evidence against them. But when taken as normative, the (mis)alignment with experimental evidence is no longer an evaluative criterion. It is descriptive models that should deliver accurate predictions. The goal of a normative model is to "deliver good advice about decisions, inferences, the structure of beliefs and the like – at some appropriate level of abstraction" (p.1347). Moreover, the author does leave the possibility open for an empirical test for normative theories (still, in a different

[26] Kasbergen (2017) looks into abstraction and idealization in EL. Among others, the author interviewed researchers in the broad EL community. They similarly remarked the importance of simplification.

sense of that of descriptive models), but also acknowledges that devising such a test is not an easy task. Lemmon and Henderson (1959) and Lemmon (1967) put forward the idea that EL does justice to *Logically Perspicuous Knowers*. While real people are not LPKs,[27] this logical fiction serves as the embodiment of the "rational man" (basically what SRP envisages). While this response is sometimes attacked as a move that disconnects logic and actual knowledge, the normative role of EL lies in modelling what we ought to know, namely what a LPK knows.

This defense still faces counterarguments: we have argued that there are good reasons to account for the fact that agents do not know *all* consequences of their knowledge even when studying how they *ought to* reason. This was precisely the gist of the counterarguments against SRT. A logically omniscient agent who knows p, also knows $p \wedge p$, $p \wedge p \wedge p$, $\ldots$. But why *should* a human agent, subject to the Finitary Predicament, waste her time and memory, as if they were infinite, to come to know all these statements? Therefore, forcing one to commit to models that are *either* non-normative *or* representing omniscient agents might well be a false dilemma. The empirical evidence reveals that mistakes are systematic and that even from a normative view, it makes sense to study the underlying reasoning processes of agents and what can be *feasibly* asked of them. The real challenge is to determine where the cutoff of reasoning lies. In our alternative picture, descriptive facts, e.g. regarding limitations of time and memory, are instrumental in determining the extent of feasibility. This is why we propose modelling how a rational agent comes to know things, informed by empirical facts to ensure that *ought* actually implies *can*.

This is not to derive *ought* from *is*. It still leaves room for evaluative judgments and thus preserves the normative status of logical modelling. When we design an exam for our students, we do not neglect their cognitive resources and situational constraints. A number theory professor who expects the students to prove Goldbach's Conjecture just because they have been taught about Peano axioms would hardly qualify for the teacher of the year award. Instead, teaching staff usually considers how much time students have to complete the exam or whether the exam is open-book (hence relieving part of the burden on memory). Similarly, it is usually expected that exams are distributed well in the exam weeks so that students will not have to split their attention to more than one subject per day. In making these adjustments, we do not necessarily proclaim that anything goes for the students' performance or that the exam serves no evaluative purpose. Even a non-evil number theory professor would expect that students solve a standard textbook exercise.

Overall, just as standard EL stands as a formalization of the standard picture of rationality, an alternative logical model is needed to serve as the formalization of the alternative picture of rationality, without forfeiting its normative status.[28]

[27] This is exactly why Hocutt (1972) criticizes EL; it fails to do justice to real knowledge, e.g. that of the *Logically Obtuse Man*.

[28] Hintikka's own understanding of the problem did not presuppose any kind of defense of

Line of defense 4: Redefining knowledge. Hocutt (1972) attacked EL by claiming that it is either not epistemic (i.e. it does not have anything to do with the notion of knowledge) or it is not a logic (i.e. it reveals no distinctive logical features). One defense against this attack is through clarifying what is meant by knowledge in epistemic logical systems. For example, we can distinguish "active" and "virtual" knowledge: people *virtually* know all consequences, but not *actively* (Hintikka, 1962, p.34), as the empirical evidence would indeed predict. This comes close to distinctions between implicit and explicit notions (Levesque, 1984). Thus, according to this defense, the empirical evidence does not threaten the notion of knowledge that epistemic logicians have in mind. But Hocutt rightfully observes that this vague distinction of notions seems to beg the question: just like LPKs, virtual knowledge seems to be constructed so that it fits exactly the principles of standard EL. Another explication of knowledge is in terms of defensibility, which was also criticized by Hocutt as being reducible to either LPK or virtual knowledge, hence inheriting the same criticisms.

2.5.2 Benefits

Given that there are cracks in the standard logical paradigm in light of the empirical evidence and its philosophical significance, one might question whether there should even be a place for (epistemic) logical modelling in the debate. Contrary to this, we argue that logic has a meaningful interaction with both epistemology and cognitive science.

For instance, van Benthem (2006, 2008c) gives a range of examples to show that despite their differences, EL and epistemology do have overlapping agendas and their interaction has been worthwhile and fruitful. More specifically, formal logical frameworks, building on or enriching the received view, have helped advance the debate underpinning theories of knowledge and provided more fine-grained notions of knowledge, belief, preference, further studying their interplay and revision. Moreover, DEL has provided the tools to study representation and action together, bringing logic together with those processes that in reality shape our mental states (van Benthem, 2008b). This has also facilitated the study of group notions of knowledge/belief, the progress of social epistemology and the analysis and resolution of epistemic paradoxes. Hendricks and Symons (2006) also list fruitful connections between logic and core epistemological issues pertaining to skepticism, belief revision, and the Rationality Debate, concluding that the interplay between logic and epistemology works dialectically. To support these claims the authors also mention how the *Dynamic Turn* in logic (van Benthem, 2004) dissolved its understanding as necessarily static and classical. This is especially important given that this understanding has largely contributed to the

his standard systems due to normativity (Hintikka, 1962, p.37). Similar arguments against the normative defense appear in Hocutt (1972) and Stalnaker (1991).

misconception about logic's role in human reasoning as an outdated and inflexible tool.[29] Finally, Klein (2015) identifies benefits of the use of formal systems, including logical ones, in approaching a given target system: (i) clarification (disambiguating properties of the target system, identifying structural patterns), (ii) verification (checking if an informal argument is in fact conclusive), and (iii) exploration (unveiling additional properties of the target systems).

There are also benefits reaped from the interaction of logic and cognitive science. Isaac et al. (2014) discuss the relationship between logic and cognitive science, with a particular focus on the role of non-monotonic logic (and complexity theory) in bridging the gaps between Marr's levels of analysis.[30] The authors discuss the contribution of logic in clarifying and comparing cognitive theories, and driving empirical predictions. The latter, upon being tested, may motivate the revision of the formal theory, thus resulting in a mutually beneficial interplay between the disciplines. van Benthem (2008a) considers the apparent discrepancy between logic and empirical evidence and explains how logic (especially DEL) can raise interesting questions in psychology of reasoning, especially with respect to long-term reflective behaviour, and provide modelling predictions that might not always be accurate but whose essential contribution is the principled generation of testable hypotheses and experimental designs, that help us better understand and clarify cognitive attitudes, their formation and interaction. Verbrugge (2009), focusing on social cognition, draws similar conclusions on the fruitful interaction between logic and psychology, and proposes new questions in this research agenda that lie in the intersection of experimental results, formal logical modelling, and computational cognitive modelling. We find applications of this combination to the study of human strategic reasoning in the Marble Drop Game, with different types of opponents ("rational" and "surprising") in Ghosh et al. (2014) and Ghosh and Verbrugge (2018), respectively. It therefore becomes clear that logic cannot only help us approach existing problems, but also identify new ones that were previously unseen or underestimated, concerning deductive reasoning, introspection, and higher-order reasoning.

[29] For example, Goldman (1978) claims that his tripartite *epistemics* project is beyond the scope of logic; for example, logic cannot account for epistemic principles informed by cognitive limitations. Contrary to this claim, DEL has the means to bridge this gap and account for more fine-grained mental state classifications, resource-bounded reasoning and its defects, which all seem to be desiderata of the *epistemics* agenda.

[30] It is noteworthy that this also touches upon the symbolic/connectionist debate regarding formalisms for cognition and suggests, perhaps surprisingly, that logic has an important role in bridging the two, surveying some attempts in this direction. Arguments in this direction are also made by van Benthem et al. (2020). Also consult Leitgeb (2008) and van Benthem et al. (2007) for collections of works on the interface of logic and cognitive science.

2.5.3 Bridging logic and empirical evidence

Empirical evidence gives us good reasons to revise SRP (Section 2.4) and its possible formalizations (Section 2.5.1). In the same way that EL provides formal tools for traditional epistemology, a new logical system may formalize the naturalized view of epistemology, which still has a normative component, this time one that values empirical facts too. Therefore, there are benefits in informing logic with empirical evidence.[31] Section 2.5.2 explored the other way round: the benefits of logical systems, especially dynamic systems, for epistemology and cognitive science alike. The challenge in fully realizing this interaction is striking a delicate balance that allows us to harvest the benefits of logic, without submitting to an isolationist and defensive view. In van Benthem's words:

> Now comes my simple declaration of faith. Logic is of course not experimental, or even theoretical, psychology, and it approaches human reasoning with purposes of its own. And a logical theory is not useless if people do not quite behave according to it. But the boundary is delicate. And I think the following should be obvious: if logical theory were totally disjoint from actual reasoning, it would be no use at all, for whatever purpose! (van Benthem, 2008a, p.2)

This calls for richer logical systems that capture finer distinctions in reasoning and study information along with the processes (inference, observation, communication) that generate it, instead of taking them for granted. Moreover, these systems should respect the extent to which these processes can be carried out. Recall that, broadly speaking, the goal is to formalize the alternative picture which asks us to extend RE with interdisciplinary empirical evidence. More concretely, we should use the DEL toolbox to capture (a) the experimental evidence on the limitations of human agents, to ensure that our normative standard is such that *ought implies can*, but also that (b) reasoning is not uniform, but it is often individuated by different types of processes, e.g. inputs of explicit reasoning often arise from automatic and implicit processes.

Of course, this task comes with its own challenging trade-off between simplification and accuracy. We have seen that standard EL tends to "underfit" human reasoning. We therefore plan to modify the possible-worlds apparatus in a way that increases its complexity and forfeits part of its simplicity to the quest for an alternative picture. However, it is important not to refrain from making simplifying assumptions altogether: not only is this task futile, for it undermines the very accessibility of a formal system, it also is undesired, for it relinquishes

[31] While we have focused on logical reasoning and formalizations, this point also applies to probabilistic modelling. In a common pattern, standard models have been criticized due to empirical findings, e.g. the *Linda problems* (Tversky and Kahneman, 1983), thus necessitating a shift to an alternative view. While probabilistic modelling falls outside the scope of this chapter, the topic is not disjoint from DEL. We will discuss this in more detail in Chapter 9.

the potential fruitfulness of simplicity (recall Section 2.5.1). Besides, we do not wish to "overfit" human reasoners either. This would compromise the alternative picture for the sake of a nihilistic one, against which we have already argued. The benefits of Section 2.5.2 would be hindered if we were to *indiscriminately* overload an alternative logical formalization so that it merely copies the behaviour of the subjects of one or another experiment.

Acknowledging the trade-off, and that formal models are subject to continuous refinement, we break down the larger task into smaller research questions. We aim at building logical systems in accord with the alternative picture explicitly representing the reasoning steps of people on (i) deduction, (ii) introspection, (iii) mental state attribution to others, and the effort they require, while also doing justice to the impact of the various types of mental processes.

These desiderata will be pursued in the remainder of the dissertation:

- In Part II, we focus on the reasoning of a single agent. In particular:
  - Chapter 3 discusses the representation of the deductive reasoning steps and the effort these require of a bounded agent.
  - Chapter 4 provides an extension of the framework of Chapter 3, suitable to represent more fine-grained propositional attitudes and incorporate the dynamics of interaction.
  - Chapter 5 builds on Chapter 4, to do justice to the impact of the dual process theories of reasoning.
- In Part III, we move to multi-agent reasoning:
  - Chapter 6 studies mental state attributions, focusing on the *formation* of beliefs about others' beliefs applied to False Belief Tasks.
  - Chapter 7 studies reasoning steps of deduction, introspection, and attribution (understood as *manipulation* of formed beliefs) in a multi-agent environment.
  - Chapter 8 examines the application of the framework of Chapter 7 to group epistemic notions and group reasoning processes.

2.5.4 Related work

There have already been attempts to design richer systems that share this motivation, to a smaller or larger extent. That is, epistemic logical systems that remedy the problems concerning unbounded (i) deductive reasoning, (ii) introspection, and (iii) reasoning about others. Most attempts have focused more on the (single-agent) problem of logical omniscience and less on higher-order reasoning idealizations. In what follows, we go through some of these.

Deductive reasoning. Let's start from the attempts aiming at the problem of logical omniscience, i.e. the idealization regarding deductive reasoning.[32]

A family of approaches is based on the inclusion of *impossible worlds* (Hintikka, 1975; Rantala, 1982a; Berto and Jago, 2019). These are "worlds which look possible and hence must be admissible as epistemic alternatives but which none the less are not logically possible" (Hintikka, 1975, p.477). When allowing semantic interpretations to also quantify over worlds that are not closed under any notion of logical consequence, the closure principles of omniscience are invalidated. However, this approach alone ignores the agents' logical competence and lacks explanatory power in terms of what really comes into play whenever we reason. In other words, it shifts to the other extreme, a trivialized view where anything goes with agents' knowledge and beliefs.[33]

D'Agostino and Floridi (2009) and D'Agostino (2010) have addressed the "enduring scandal of deduction". Deductive inference is usually regarded to be "uninformative" because the information of the conclusion is already "contained" in the premises. This is in conflict with the empirical fact that deductive reasoning is often a non-trivial and illuminating task. The authors identify the introduction of "virtual information" (assumptions that are used and then discharged) as instrumental in the resolution of deductive problems. Yet this information is not contained in the actual information we hold. With this in mind, the authors propose a hierarchy of tractable logics that represent different depths of Boolean reasoning. They suggest that these (and not classical propositional logic) should be taken as the basis of epistemic logics, thereby avoiding logical omniscience.

Another family of approaches discerns implicit and explicit attitudes, where the latter, unlike the former, are omniscience-free. For example, Levesque (1984) suggests that closure principles do not refer to what we *actually* know or believe but rather to another concept: what is implicit in what we know or believe, even without us realizing it. Formally, the distinction is usually achieved through a "syntactic" filter that gives rise to the omniscience-free explicit notions. For example, settings in the *Justification Logic* tradition (Baltag et al., 2014; Artemov and Fitting, 2020) appeal to notions of justification and evidence in order to tell apart implicit and explicit attitudes. These additional ingredients help in avoiding problematic closure principles and they come with their own operations that can be seen as corresponding to human mental operations. In Fagin and Halpern (1987), agents have to be additionally *aware* of something to know it explicitly. The models are augmented by an *awareness function*, yielding the formulas the agent is aware of. However, forms of the problem may persist, semantic and syntactic notions become conflated, and it is not clear how logical competence or

[32] For other overviews, see Sim (1997); Moreno (1998).

[33] Cozic (2006) advocates for the use of such worlds from a decision-theoretic viewpoint, making use of probabilistic structures. The challenge mentioned above also applies to this view.

resource-boundedness could be accounted for.[34]

These remarks also apply to the *local-reasoning structures*, appearing in Fagin and Halpern (1987); these account for an agent holding inconsistent knowledge/beliefs by viewing her as a "society of minds", formally captured via sets of sets of worlds each corresponding to the set of worlds the agent considers possible in a given frame of mind. Hawke et al. (2019) address the problem of logical omniscience by adding the *topic* of sentences into the picture. This tackles one dimension of the problem, which is nonetheless different from the boundedness of real reasoners. The importance of thematic connections also appears in Hoek (2020), where the author treats (minimally rational) beliefs as answers to questions and revisits deductive reasoning as an issue of asking new and good questions. The variety of reasons why people fall short of logical omniscience is identified by Égré (2020), who surveys approaches to address the problem. Interestingly, the author observes the importance of the dynamics of belief formation in addressing logical omniscience as this is not accounted for in some approaches, e.g. aiming at belief fragmentation as a source of fallibility.

Moreover, another family of approaches uses *Montague-Scott semantics*. This substitutes the standard relational structures with *neighborhood structures* – a set *of sets* of worlds is assigned to each world by a neighborhood function. The modal truth clause is modified accordingly, yet *Closure under Logical Equivalence* eventually persists (Fagin et al., 1995, Chapter 9). The implicit/explicit distinction under this type of semantics appears in Velázquez-Quesada (2013) and Fernández-Fernández and Velázquez-Quesada (2019).

Closer to our reasoning-oriented understanding of the problem, there are attempts formally accounting for the reasoning steps or the time it takes to derive logical consequences (Duc, 1997; Elgot-Drapkin et al., 1999). In a similar spirit, Alechina and Logan (2009), Jago (2009), and Ågotnes and Alechina (2006) use temporal-like state-transitions corresponding to inference rules. The deduction model of Konolige (1986) uses *belief sets* closed under an (incomplete) set of inference rules (so weaker closure still applies). Combinations of the aforementioned ideas appear in Velázquez-Quesada (2011), where awareness sets can be modified depending on the agent's applications of inference rules, and in Bjerring and Skipper (2018), using operators standing for a *number* of reasoning steps, and impossible worlds.

However, we believe that it might be better to abstain from a generic notion of reasoning process (be it number of steps, arbitrary rule applications, or time intervals) and not presuppose the existence of an arbitrary cutoff on reasoning. Instead, we should introduce *explicitly* applications of different rules, their chro-

[34] A recent exception is Lorini (2020), whose central notion is that of a *belief base*, a set that contains the explicit beliefs of an agent, and is not necessarily logically closed. The notions of possible world and epistemic/doxastic accessibility are derived from that, instead of being taken as primitives. It would be interesting to incorporate the dynamics of reasoning and resource-boundedness in such logics of implicit and explicit beliefs.

nology, and the cognitive effort they require. In this way, we can exploit studies in psychology of reasoning – that usually study *individual* inference rules in terms of cognitive difficulty – and thus bridge the logical approach with empirical facts.[35]

Introspection. Williamson proposes a system of *inexact knowledge* that drops both positive and negative introspection altogether (Williamson, 1992, 2000). Bonnay and Égré (2009, 2011) identify shortcomings in this analysis and propose an alternative framework for inexact knowledge. In particular, their token semantics determines to what degree of modal depth introspection goes and how this depends crucially on one's available resources. However, in accord with arguments given above, it may be beneficial to specify exactly what these resources are, and to take into account the depletion of resources induced by the dynamic nature of introspection. Liu (2009) identifies several sources of agent diversity, such as introspective abilities, observation powers, memory capacity, and revision policies, and brings different types of agents (e.g. introspective and non-introspective) together in a multi-agent setting. Still, even a single agent might become more or less capable of introspection, depending on the availability of resources or the context in which her beliefs arise. It is interesting to study the evolution of the agent's reasoning both against her own potential and in interaction with different kinds of agents. Jago (2009) extends his rule-based approach on EL with introspective rules, that allow agents to gain introspective beliefs. This is captured in the context of temporal-style transitions and not of DEL, which might be better equipped to capture how introspective actions induce a transformation, and importantly, a resource-consuming one. Fervari and Velázquez-Quesada (2019) view introspection not as a property but as an *action*; they provide actions of *general* introspection, *particular* introspection, and *one-step* introspection, which modify the accessibility relations. The motivation is similar to ours: aiming at realistic agents who need to take action to introspect. In our view, the most adequate notion is that of *one-step* actions, since we need to account for *individual* steps – actions that once taken do not guarantee full introspection. This seems to be more in line with importing empirical evidence on introspective effort.

Reasoning about others. It is worth noticing that there have been logical formalizations that incorporate findings on ToM. We have already mentioned the work of Ghosh et al. (2014) and Ghosh and Verbrugge (2018), dealing with human resource-bounded reasoning in games. The authors use logic to formulate different reasoning strategies, and then compare cognitive models based on them

[35] Recall Section 2.3. Some concrete examples: Rips (1994) claims that the difficulty of a complex reasoning tasks depends on the length and the difficulty of the rules involved in the mental proof constructed for it. The experiments of Rijmen and De Boeck (2001) verify that different rules require different weights, and the authors claim that associating difficulty with the number of mental models (Johnson-Laird et al., 1992) performs relatively well, but it still is a rather coarse grained measure to capture the variance; in Zhai et al. (2015), empirically calculated weights are attached to different rules.

with actual human performance on basis of reaction times. This allows for the construction of cognitive computational models with similar task performance to that of the actual players. Other examples include Stenning and van Lambalgen (2008), providing a non-monotonic closed-world reasoning formalization of first-order False Belief Tasks, implemented within logic programming; van Ditmarsch and Labuschagne (2007), focusing on three case studies of agents with special higher-order reasoning properties; Bolander (2018), using a variant of DEL action models to model False Belief Tasks, later implemented on a humanoid robot (Dissing and Bolander, 2020); and Braüner (2015); Braüner et al. (2016), offering a hybrid-logical study of the tasks. These frameworks bring ToM literature together with logic, but focusing on *the formation of beliefs*, e.g. due to observation, communication, etc. It would be interesting to investigate this belief formation, beyond specific types of agents, and to also explain why higher-order reasoning eventually halts due to cognitive fatigue. Another dimension of bounded reasoning about others is that of *manipulation of beliefs*, which we discuss next.

In the work of Alechina et al. (2008), the agents' beliefs are essentially given by a set that need not be logically closed, thereby avoiding omniscience. The authors also introduce certain actions of inference (*resolution*), introspection, others' resolution, and others' introspection as well as bounds on memory, the level of nesting of higher-order beliefs, and communication. It would be interesting to consider how the already established progress of DEL could help in incorporating empirical evidence in a similar setting that deals with bounded chains of deductive reasoning steps, introspective steps, and attributions of reasoning steps to others.

Balbiani et al. (2019) develop a multi-agent framework for non-omniscient, resource-bounded agents who can reason and enrich their beliefs through actions of perception and inference. A crucial feature is a distinction of *background knowledge* and *explicit beliefs*, which is seen on a par with a distinction between long-term memory and working memory. The semantics also mirrors this distinction as it employs a combination of relational semantics (for background knowledge) and neighborhood semantics (for explicit beliefs). As a result, the former is subject to closure properties of logical omniscience, while the latter is not, save for closure under logical equivalence. Moreover, the actions modify appropriately the neighborhood function and the accessibility relations. Still, it is important not to assume any omniscient notion and to explicitly add cognitive parameters in the models, e.g. to capture the differential contributions of rules, as rules do not impose equal cognitive burden. Also, neighborhood semantics maintains closure under equivalence, which is important if we consider framing effects whose study is central to investigations of human reasoning. Moreover, tools from DEL, like plausibility action models (Baltag and Smets, 2008b), might be more flexible in that they capture more fine-grained definitions of attitudes and actions, but also other notions like *preferences* and their upgrades (van Benthem and Liu, 2007).

2.6 Summary

Standard Epistemic Logic was seen as the formalization of the Standard Picture of Rationality. Given the rich empirical literature on human reasoning, we also discussed experimental results on the reasoning performance of real people. This diverges from the norms of the Standard Picture and as such attacks the Standard Rationality Thesis. We examined responses given by proponents of the Thesis, in order to restore its credibility, despite these findings. We argued against these responses and concluded with the need for an alternative rationality picture, wherein descriptive facts, on cognitive limitations and the variety of mental processes, have a key role. These results, i.e. the empirical evidence and its repercussions in replacing the Standard Picture, hinted at the need for another logical formalization, now corresponding to the alternative picture. Yet, we still considered whether there are independent reasons to stick to the **S5** (/**KD45**) modelling for knowledge(/beliefs), despite the shift from one epistemological picture to another. We concluded that it is indeed worthwhile to seek alternative logical systems that properly bridge formal modelling and psychology of reasoning, to the mutual benefit of the disciplines studying human reasoning.

Part II

Single-agent Reasoning

Chapter 3

A dynamic epistemic logic for a resource-bounded agent

In this chapter, we devise an alternative framework that aims at amending the relational semantics of EL towards resolving the problem of logical omniscience.[1] The framework builds on impossible-worlds semantics, but it nonetheless avoids the extreme of logically incompetent agents. That is, we want to capture how resource-boundedness affects the reasoning processes underpinning knowledge acquisition of *competent* agents. This is why we also make use of the DEL toolbox to keep track of the reasoning steps available to the agent, their orderly applications, and the cognitive effort they require.

More specifically, our model is a variant of Kripke models extended with impossible worlds and quantitative components capturing the agent's cognitive capacity and the cognitive costs of rules with respect to certain resources (e.g. memory, time). The model updates capture the agent's applications of deductive inference rules and modify its components (epistemic accessibility, cognitive capacity) according to each application's effect. We further provide a sound and complete axiomatization, through a method that connects this semantic approach to logical omniscience with more syntactically-oriented ones, and thus allows for comparative remarks between the two.

The chapter is structured as follows: in Section 3.1 we unfold the background that we will rely on to devise the framework. This is in turn presented in Section 3.2. The method that allows us to extract a sound and complete axiomatization is given in Section 3.3.

3.1 Background

We look into important background notions, namely impossible-worlds semantics and DEL, and attempts against logical omniscience that have made use of them.

[1] The chapter is based on Solaki (2019); Smets and Solaki (2018); Solaki and Smets (2021).

3.1.1 Impossible worlds

One proposal against the problem of logical omniscience has been to include impossible worlds (Hintikka, 1975): worlds that *look* possible to the agent but are not *logically* possible. A proposal along these lines is given by Rantala (1982a) whose models include both possible (*normal*) and impossible (*non-normal*) worlds; the truth values in the latter are given *directly*, and not recursively as in the former. Provided that a propositional attitude is interpreted by quantifying over impossible worlds too, the closure properties that give rise to logical omniscience are broken. In a sense, one can view them as witnesses of the agent's fallibility as any non-ideal agent might entertain an inconsistent or incomplete scenario. This attempt is extended to quantified modal logics by Rantala (1982b).

The unrestricted use of such logically "anarchic" worlds has been criticized because it generally yields trivialized systems, wherein anything goes with an agent's reasoning (Williamson, 2020).[2] Granted, we avoid the problem of logical omniscience, but only because we resort to the other extreme: logical incompetence. As a result, the inclusion of such worlds does not suffice to address the challenges of modelling deductive reasoning explained in Chapter 2. Jago (2006) and Berto and Jago (2019) similarly observe that this approach collapses to any approach whereby knowledge or beliefs are given by an arbitrary set of formulas.[3]

However, the problem of demarcating which impossible worlds should be accessible to non-omniscient but competent agents is not a mere defect of this particular choice of semantics. Rather, it is illustrative of the broader challenge of finding an alternative picture of rationality that avoids the unattainable requirements of the standard picture without submitting to a nihilistic view where anything goes. This challenge is reflected on impossible-worlds semantics – and perhaps on other modifications of Kripke semantics against logical omniscience. We have to allow for impossible worlds to avoid omniscience, but we have to disallow for "trivially" impossible worlds to accommodate logical competence. The quest for this delicate balance is also evident in Jago (2013)'s discussion of the "problem of rational knowledge" and Bjerring (2013)'s impossibility result.

To anticipate the design of our approach to impossible-worlds semantics, we have to recall the desiderata of our alternative rationality picture (Chapter 2). These, however, cannot be realized in a purely static framework, whose agents are stuck either in an idealized or in an incompetent state. To do justice to agents who are fallible, but still try to refine their knowledge, we have to study their *dynamic* reasoning processes. This naturally brings us to the next prerequisite.

[2] Another objection against this type of semantics that might be relevant here is the "compositionality objection", which has been addressed by (Berto and Jago, 2019, Chapter 8).

[3] One way to impose *some* degree of logical structure in models with impossible worlds is to accept that they are closed under *some* closure principles, albeit not the ones of classical logic. Still, this is inadequate. Closure under another, weaker logic is equally worrisome for the purpose of modelling bounded agents. Consider, for instance, a paraconsistent logic; we should not expect that a finite agent knows all *paraconsistent* consequences of her knowledge.

3.1.2 Dynamic Epistemic Logic

In reality, our mental states change in the aftermath of certain actions. This is why static languages are sometimes supplemented by dynamic operators that stand for informational actions. These induce *model transformations*. They transform a model representing the epistemic state of an agent to a new one, that represents the updated epistemic state. Given action α, a formula of the form $\langle\alpha\rangle\phi$, where $\langle\alpha\rangle$ is a dynamic operator, is evaluated in a model by evaluating what the truth value of ϕ is at the transformed model, that resulted from carrying out α. A common example is the logic of *public announcements* (Plaza, 2007), (van Ditmarsch et al., 2007, Chapter 4). What is of special interest for our purposes is how DEL can help us track an agent's underlying reasoning processes. This is evident in some of the attempts surveyed in Section 2.5.4. Below, we focus on one attempt along these lines that involves both DEL and impossible-worlds semantics.

A combination of impossible-worlds semantics and DEL is given by Bjerring and Skipper (2018), building on ideas of Rasmussen (2015). The authors provide a doxastic framework that uses dynamic operators ($\langle n\rangle$) standing for a number of reasoning steps (applications of inference rules). The corresponding action induces model transformations; these, roughly, reflect the variations of worlds the agent can come to consider possible, by performing a chain of reasoning of the given length. This combination is useful for it avoids omniscience (due to the impossible worlds) but it simultaneously does justice to the intuition that competent agents can come to believe certain (but not all) consequences of their beliefs. There is, however, room for improvement. To begin with, it is not clear how the length of the reasoning process (the number n) can be independently motivated and on what empirical grounds it can be based. In the absence of a plausible method, such as an empirical indication, it is natural to wonder what differentiates n and $n+1$ steps in picking out a chain of reasoning as "too long" and thus what renders a belief unattainable. Inevitably, questions of vagueness discussed by Jago (2014) emerge again, although the authors argue against it. In addition, it is not clear how this proposal would deal with agents holding inconsistent beliefs.

In what follows, we design a dynamic, impossible-worlds semantics that spells out a bounded agent's reasoning processes. However, it does not presuppose an arbitrary number of reasoning steps and it allows us to practically witness to which extent they evolve: to the extent allowed by the agent's resources.

3.2 Resource-bounded deductive reasoning

3.2.1 Syntax

To begin with, we need a logical language where the agent's steps of deductive reasoning are explicitly introduced, since we do not want to take them for granted, swept under one unique knowledge modality.[4] We view these steps as applications of inference rules. To that end, we first define:

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

We denote the set of premises and the conclusion of ρ by $pr(\rho)$ and $con(\rho)$ respectively, while $\mathcal{L}_R$ denotes the set of all inference rules. For example, an instance of *Disjunctive Syllogism* $DS := \{p \vee q, \neg p\} \rightsquigarrow q$ is such an inference rule where $pr(DS) = \{p \vee q, \neg p\}$ and $con(DS) = q$.

Notice that rules are intended to be read as *instances* and not as *schemes* (i.e. using meta-variables which are meant to be replaced with formulas). This is because the former reading is better suited to embed cognitively plausible features, pertaining to the cognitive cost that rule-applications impose on human subjects. As we will see, there are good empirical reasons to study *individual* instances in terms of cognitive difficulty.[5] We now move to another prerequisite notion:

3.2.2. Definition (Terms). The set of terms T is defined as $T := \{c_\rho \mid \rho \in \mathcal{L}_R\} \cup \{cp\}$. It contains elements for (i) the cognitive costs of rule-applications (of the form c_ρ), and (ii) the cognitive capacity of the agent (of the form cp).

With these in place, we can proceed to the definition of our language:

3.2.3. Definition (Language). The epistemic language $\mathcal{L}_K$ is built as follows:

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

where $z_1, \ldots, z_n \in \mathbb{Z}$, $z \in \mathbb{Z}^r$, $s_1, \ldots, s_n \in T$, $p \in \Phi$, and $\rho \in \mathcal{L}_R$.[6]

The additional components compared to the standard epistemic language are:

- Quantitative comparisons introduced to deal with cognitive effort (e.g. of the form $s_1 \geq s_2$). This is possible because the terms of T essentially express the cognitive costs of inference rules and the agent's cognitive capacity.
- An operator A, introduced to capture the agent's availability of inference rules. Specifically, $A\rho$ is to say that ρ is acknowledged as truth-preserving by the agent.[7]

[4] We here focus on an *epistemic* framework, but a doxastic one can be designed similarly.

[5] In doing so, we follow the rationale of (Velázquez-Quesada, 2011, Chapter 2): rules are supposed to be applied in a generalized Modus Ponens way: if you know the premises, you may come to know the conclusion.

[6] The choice of the number r will be made precise in the next subsection. In $\mathcal{L}_K$, formulas involving $\leq$, $=$, $-$, $\vee$, $\rightarrow$ can be defined as usual in terms of the rest. This is why a formula of the form $s_1 \geq s_2$ is well-formed: it abbreviates $s_1 + (-1)s_2 \geq \overline{0}$.

[7] Operators in this spirit appear in Elgot-Drapkin et al. (1999); Ågotnes and Walicki (2004); Velázquez-Quesada (2011).

- Dynamic operators labelled by inference rules, of the form $\langle\rho\rangle$, such that: $\langle\rho\rangle\phi$ stands for "after an application of ρ, ϕ is true".[8]

Examples of formulas. The formula $(cp \geq c_\rho) \wedge A\rho$ says that (i) the cognitive capacity of the agent (to which the term cp corresponds) is greater or equal than the cognitive cost of a rule ρ (to which the term c_ρ corresponds), and (ii) the agent has rule ρ available. A formula like $\langle\rho\rangle K\phi$ is to say that after the agent manages to apply ρ, she comes to know that ϕ.

3.2.2 Defining resource-sensitive models

Our models supplement Kripke models with impossible worlds and cognitive components. Impossible worlds are employed, as in Section 3.1, to do justice to the fallibility of agents as real people might entertain inconsistent/incomplete scenarios. The novelty in our framework, as we shall see, is that the agent can gradually eliminate some of them, by taking reasoning steps to the extent they can cognitively afford them.

Still, we want to build a model aligned with the alternative picture of rationality. We should thus respect the *Minimal Rationality* of agents (recall Section 2.4.1). According to Cherniak (1986), we need a "theory of feasible inferences" where the difficulty of deductive reasoning is responsible for the agent performing *some*, but not all appropriate inferences, so in fact, we need a "well-ordering of inferences" in terms of difficulty. It is natural to connect this with the consumption of cognitive resources and use it to determine where the cutoff of an inferential chain lies. To start with, since Minimal Rationality also translates to the ability to eliminate inconsistencies, we will rule out what is an obvious case of inconsistency for any logically competent agent: explicit contradictions.

In order to describe the other components, we first need to parameterize our models by Res, denoting the set of resources (*memory*, *time*, *attention*, etc.) we want to consider. Then $r := |Res|$ is the number of these resources. Another parameter concerns the cognitive effort of agents. The cost function $c : \mathcal{L}_R \to \mathbb{N}^r$ assigns a *cognitive cost* to each rule with respect to each resource. That is, cost is a vector, as in Alechina et al. (2009a), indicating the units consumed per resource for the several reasoning steps. For example, given the resources of time and memory, and an instance of Modus Tollens (MT), $c(MT) = (3, 2)$ says that this instance consumes 3 units of time and 2 units of memory.

With the above fixed, we can introduce a *cognitive capacity* model component to capture the agents' available power with respect to each resource. Unlike the parameters, capacity will not be fixed, as resources are depleted while reasoning

[8] Operators in this spirit appear in Velázquez-Quesada (2011). The dual $[\rho]$ can be defined as usual in Modal Logic.

evolves, and it may change as a result of deductive actions.[9]

Before we give the formal definition of our models, we elaborate on the assignment of costs and capacity with respect to cognitive resources. The concrete assignments rely on empirical research. We here follow a numerical approach because the cognitive difficulty of reasoning tasks has often been explained in terms of the number and the kind of rules that have to be applied (Johnson-Laird et al., 1992; Rips, 1994; Rijmen and De Boeck, 2001). This stems from experimental observations on the different response times of people in questions involving applications of different inference rules (Marcus and Rips, 1979; Rips, 1994), limits on (working) memory capacity (Miller, 1956; Cowan, 2001), and attention (Kahneman and Beatty, 1967; Xu and Chun, 2009).

For example, an application of a rule consumes the limited time an agent can devote to a task, and the premises used in applying the rule consume elements of the limited memory. Moreover, the logical complexity of the premises might also play a role in the differential costs of inference rules. Research on the cognitive difficulty of Boolean concepts and the debates ensuing from it are especially relevant for that (Feldman, 2000, 2003; Vigo, 2006; Goodwin and Johnson-Laird, 2013). The different schools, e.g. *Mental Logic* and *Mental Models* (recall Section 2.3) interpret experimental results differently, pointing at different "measures" for the difficulty of deductive tasks (the number of steps in a mental proof, the number of mental models that have to be checked, etc.). Still, the very observation that not all inferences require equal effort is generally accepted. Since the expected difficulty of a rule (both as a scheme and as a particular instance) is not settled in the empirical realm, we have not committed to a particular view on cost assignments. Instead, we focus on providing the machinery to embed such features in formal logical modelling, regardless of one's preferred analysis.

We can now give the definition of *resource-sensitive models*, based on the above.

3.2.4. Definition (Resource-sensitive model). Given a set of r-many resources Res and a cost function c, a *resource-sensitive model* (RSM) is a tuple $M = \langle W^P, W^I, f, V_P, V_I, R, cp\rangle$ where:

- W^P, W^I are non-empty sets of possible and impossible worlds respectively. Take $W := W^P \cup W^I$.

- $f : W \to \mathcal{P}(W)$ is a function mapping each world to its set of epistemically accessible worlds.

- $V_P : W^P \to \mathcal{P}(\Phi)$ is a valuation function assigning to each *possible* world, the propositional *atoms* that are true there.

[9] 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).

- $V_I : W^I \to \mathcal{P}(\mathcal{L}_K)$ is a valuation function assigning to each *impossible* world, the formulas (*atomic or complex*) that are true there.
- $R : W^P \to \mathcal{P}(\mathcal{L}_R)$ is a function yielding the rules the agent has available (i.e. has acknowledged as truth-preserving) at each possible world.
- cp denotes the agent's cognitive capacity, i.e. $cp \in \mathbb{Z}^r$, intuitively standing for what the agent can afford with respect to each resource.

Each RSM comes parameterized by Res and c, yet we will not explicitly write them down as components of the model. This is to serve simplicity of notation but also to emphasize that these, unlike cp, are not meant to be modified in the aftermath of reasoning actions. A pair (M, w) consisting of a model M and a world $w \in W^P$, designated as the actual one from the perspective of the modeller, is called a *pointed model*. In accordance with the remarks made above, we impose certain conditions on the model:

(1) **Minimal Consistency**: $\{\phi, \neg\phi\} \not\subseteq V_I(w)$ for all $w \in W^I, \phi \in \mathcal{L}_K$

It is common in EL to ask that epistemic accessibility is reflexive, symmetric, and transitive, properties that correspond to properties of knowledge: factivity, positive and negative introspection. Still, for reasons described in Chapter 2, we abstain from assuming unlimited introspection, thus from asking that accessibility is symmetric and transitive. In the context of resource-bounded agents, it is reasonable to extend considerations of non-ideal performance to higher-order reasoning as well. This point will be addressed in detail in Part III. In what follows, we impose reflexivity and thus factivity of knowledge:

(2) **Reflexivity**: $w \in f(w)$, for all $w \in W^P$

Because of factivity, we also need to ensure that the inferences refining the agent's epistemic state are truth-preserving.[10] To that end, and assuming that propositional formulas are evaluated as usual at possible worlds, we ask for:

(3) **Soundness of rules**: for $w \in W^P$, if $\rho \in R(w)$ then $M, w \models tr(\rho)$, where
$$tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \to con(\rho)$$

To capture what has been called "coherence for rules" (Velázquez-Quesada, 2011, Chapter 2), we also ask that the rules available to the agent are preserved by epistemic accessibility.[11]

(4) **Rule-availability**: for $w, u \in W^P$, if $u \in f(w)$ and $\rho \in R(w)$ then $\rho \in R(u)$

[10] For a doxastic framework, this requirement could be relaxed, for beliefs could be developed through non-truth-preserving inferences as well.

[11] We will study actions that modify the availability of rules while still respecting the condition in Chapter 4.

3.2.3 Logical dynamics: applying a rule

The language contains operators for actions capable of changing an agent's epistemic state following a certain rule-application. The semantic effect of these actions is, as usual in DEL, captured via model transformations. One way to capture the change induced by applications of inference rules is to encode them on the structure of our models. The effect of applying a rule is, intuitively, an expansion of the agent's information. As a result, we have to impose an additional model condition to ensure that there are worlds capable of representing such expansions.

(5) **Succession:** For every $w \in W^I$, if: (a) $pr(\rho) \subseteq V_I(w)$, (b) $\neg con(\rho) \notin V_I(w)$, and (c) $con(\rho) \neq \neg\phi$ for all $\phi \in V_I(w)$, then there is some $u \in W^I$ such that $V_I(u) = V_I(w) \cup \{con(\rho)\}$.[12]

We call u a *ρ-expansion* of w. Regarding possible worlds, for $w \in W^P$, its *ρ-expansion* is taken to be w itself, because of its deductive closure.

Succession is in fact reminiscent of the *Comprehension Principle* adopted by Bjerring and Skipper (2018) (cf. (Nolan, 1997)). According to it, the model is rich enough to contain an impossible world for any incomplete/inconsistent set of formulas. Comprehension is adopted to ensure that the model represents *all* the different ways the world could *not* possibly be. Succession is similar to, albeit more moderate than, Comprehension. It asks that the model is rich enough to represent the deductive reasoning of the agent, without necessarily committing to a view on the metaphysical status of impossible worlds.[13]

We name the class of RSMs fulfilling (1), (2), (3), (4), and (5) *epistemic* RSMs. Next, we define the *ρ-radius*, in order to represent how a rule ρ triggers an informational change, to the extent that Minimal Consistency is respected.

3.2.5. Definition (ρ-radius). The ρ-radius of a world $w \in W^P$ is $w^\rho := \{w\}$. The ρ-radius of a world $w \in W^I$ is given by:

$$w^\rho := \begin{cases} \{w\}, \text{ if } pr(\rho) \nsubseteq V_I(w) \\ \varnothing, \text{ if } pr(\rho) \subseteq V_I(w) \text{ and} \\ (\neg con(\rho) \in V_I(w) \text{ or } con(\rho) = \neg\phi \text{ for some } \phi \in V_I(w)) \\ \{u \mid u \text{ is a } \rho\text{-expansion of } w\}, \text{ if } pr(\rho) \subseteq V_I(w) \text{ and} \\ \neg con(\rho) \notin V_I(w) \text{ and } con(\rho) \neq \neg\phi \text{ for all } \phi \in V_I(w) \end{cases}$$

A rule whose premises are not true at an impossible world does not trigger any change; this is why the only expansion is the world itself. A rule that leads

[12] Note that $=$ between formulas stands for syntactic identity.

[13] We will later design an alternative way to represent this (Part III). The alternative relies less on Comprehension-like principles and more on a technical combination of impossible-worlds models and the so-called *action models* of DEL.

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. As ρ-expansions expand the state from which they originate, inferences are not defeated as reasoning steps are taken, hence Succession warrants monotonicity, but to the extent that Minimal Consistency is respected. Notice that w's ρ-radius amounts to $\{w\}$ for $w \in W^P$, due to the closure of possible worlds, while the radius of an impossible world contains another impossible world. The radius is instrumental in modifying epistemic accessibility in the transformed model, after a rule-application.

3.2.6. Definition (Model transformation by application of a rule ρ). Take RSM $M = \langle W^P, W^I, f, V_P, V_I, R, cp\rangle$. The transformation of M by an application of ρ is the model $M^\rho := \langle W^P, W^I, f^\rho, V_P, V_I, R, cp^\rho\rangle$ with:

$$f^\rho(w) = \bigcup_{u \in f(w)} u^\rho \qquad\qquad cp^\rho = cp - c(\rho)$$

That is, given a pointed model (M, w), the updated pointed model (M^ρ, w) is such that (a) w's epistemically accessible worlds in M are replaced by the elements of their ρ-radii, and (b) the cognitive capacity is reduced by the cost of performing the ρ-step. It is easy to check that the conditions on epistemic RSMs (1-5) are preserved through this operation.

3.2.4 Truth clauses

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.

3.2.7. Definition (Term interpretation). Given $M = \langle W^P, W^I, f, V_P, V_I, R, cp\rangle$ parameterized by resources Res and the cognitive cost function c, 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 truth 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. Therefore, the truth clauses are given by:

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

$$\begin{array}{ll}
M,w \models p & \text{iff } p \in V_P(w) \\
M,w \models z_1 s_1 + \ldots + z_n s_n \geq z & \text{iff } z_1 s_1^M + \ldots + z_n s_n^M \geq z \\
M,w \models \neg\phi & \text{iff } M,w \not\models \phi \\
M,w \models \phi \wedge \psi & \text{iff } M,w \models \phi \text{ and } M,w \models \psi \\
M,w \models A\rho & \text{iff } \rho \in R(w) \\
M,w \models K\phi & \text{iff } M,u \models \phi \text{ for all } u \in f(w) \\
M,w \models \langle\rho\rangle\phi & \text{iff } M,w \models cp \geq c_\rho,\ M,w \models A\rho \text{ and } M^\rho,w \models \phi
\end{array}$$

We say that a formula is *valid in a* RSM M if it is true at all *possible* worlds of the model (as usual in impossible-worlds semantics), and *valid in the class of epistemic* RSMs (or simply *valid*) if it is valid in any epistemic RSM M. 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 reasoning step in order to gradually minimize her ignorance.

3.2.5 Discussion

Given our clause for K, the presence of impossible worlds, where formulas are assigned a truth value directly rather than recursively, suffices to break the closure principles of logical omniscience. This is clear in the following invalidity:

$$\not\models \bigwedge_{\phi \in pr(\rho)} K\phi \to Kcon(\rho)$$

On the other hand, despite being fallible, an agent can still come to know consequences of her knowledge and gradually eliminate impossibilities she initially entertained. The framework is *dynamic*, like real reasoning is. Consider the truth conditions for epistemic assertions like $K\phi$ prefixed by a rule ρ; they require that (a) the rule is cognitively affordable, (b) the rule is available to the agent, (c) ϕ follows from the accessible worlds via an application of ρ. Cognitive capacity, decreasing suitably after every rule-application, determines to which extent consequences of one's knowledge can come to be in turn known. This cutoff is therefore cognitively informed and not arbitrarily fixed. 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. This is clear in the following validity:

$$\models \bigwedge_{\phi \in pr(\rho)} K\phi \wedge A\rho \wedge (cp \geq c_\rho) \to \langle \rho \rangle K con(\rho)$$

Notice that while we do use impossible worlds (as described in Section 3.1), ours is a *dynamic* framework and "impossibilities" entertained by agents are restrained by the ability to perform bounded reasoning. This restores explanatory power and avoids a trivialized view where anything goes. This idea is shared by Bjerring and Skipper (2018). Notice, though, that spelling out the individual rules and the effort they require is necessary to avoid fixing an arbitrary n only to stumble across the usual dilemmas involved in impossible-worlds semantics between idealization and competence. This elaborate specification is crucial in bridging epistemic frameworks with empirical facts on the difficulty of individual reasoning steps. It is along with these lines that the attempt differs from others utilizing DEL in resolving logical omniscience (see Section 2.5.4). 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 a response to the problem of logical omniscience.

We now provide an example, inspired by the abstract selection task, to see this type of modelling in action:

3.2.9. Example. Consider the following scenario: agent Alice is given 2 cards, each has a number on one side and a letter on the other. Alice knows that if a card has a *v*owel on one side, it has an *e*ven number on the other. Suppose that the 1st card has 'A' on its visible side (denoted by v_1 to say that "the first card has a *v*owel"), and the 2nd card has '7' (denoted by $\neg e_2$; e_2 stands for "the second card has an *e*ven number"). Alice also has the following rules available, $MP := \{v_1 \to e_1, v_1\} \rightsquigarrow e_1$ and $MT := \{v_2 \to e_2, \neg e_2\} \rightsquigarrow \neg v_2$. Fixing *time* and *memory* as our resources of interest, take $cp = (5, 4)$, $c(MP) = (1, 2)$ and $c(MT) = (3, 2)$.[14]

The initial (pointed) model for Alice, who entertains an incomplete and an inconsistent world, is called M and it is depicted in Figure 3.1 (left). Clearly, $M, w \models K(v_1 \to e_1) \wedge Kv_1 \wedge K(v_2 \to e_2) \wedge K(\neg e_2)$ but $M, w \not\models Ke_1$ and $M, w \not\models K(\neg v_2)$. Alice has not *automatically* unpacked the consequences of her knowledge.

[14] In Section 3.2.2, we argued that proper assignments depend on empirical research and provided some relevant sources. The choices in this example merely illustrate the workings of the framework and are motivated as follows. On capacity: the memory value is due to the bounds of working memory described in Cowan (2001) and time is indicative of a fixed interval in which participants should complete a reasoning task. On costs: both rules consume the same units of memory (determined by the number of premises alone, since Boolean complexity does not differ markedly for the instances in question), while the difference in time can be explained in terms of the asymmetry between MP and MT, presented in Section 2.3.

But what should Alice derive? In seeing the vowel card and performing a MP step (since this is available and affordable), she comes to know e_1. Then, in seeing the number card and performing a MT step (since this is available and affordable), she comes to know $\neg v_2$. A depiction of the pointed MP, MT-updated model is given in Figure 3.1 (right). The initially accessible impossible worlds have been replaced by the elements of their respective radii. In particular, (a) u_1 has been first replaced by its MP-radius v_1, and subsequently v_1 by its MT-radius z_1, and (b) u_2 has been replaced by its MP-radius v_3, and then eliminated from epistemic accessibility because the MT-radius is empty (i.e. the world is uncovered to be inconsistent by a MT-application). The updates, in accord with Definition 3.2.6, induce not only a change in accessibility but also a decrease in the agent's capacity by the costs of MP and MT, i.e. $cp' = (1, 0)$.

In Figure 3.1, we draw impossible worlds as rectangles and write down all formulas, atomic or complex, true there, to distinguish them from the possible worlds where we only write the atoms that are true there. For example, in the possible world w, the following are true: v_1, e_1 (thus $\neg e_2, \neg v_2$ are also true as possible worlds are maximal consistent alternatives). The arrows indicate epistemic accessibility for Alice while the dashed arrows indicate (non-trivial) rule expansions. The reflexive arrows are omitted for simplicity and the thicker node denotes the actual world. The "faded" part of the MP, MT-updated model is to show the replacement of accessible worlds carried out because of the rule-applications.

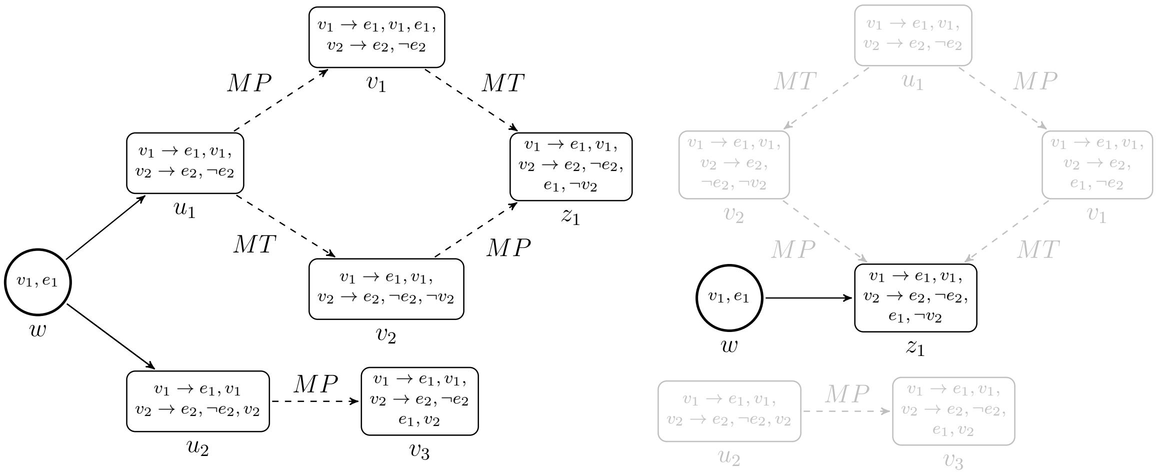

$v_1 \to e_1, v_1, e_1,$ $v_2 \to e_2, \neg e_2$
v_1
MP
MT
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2$
u_1
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2,$ $e_1, \neg v_2$
z_1
MT
MP
v_1, e_1
w
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2, \neg v_2$
v_2
$v_1 \to e_1, v_1$ $v_2 \to e_2, \neg e_2, v_2$
u_2
MP
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2$ e_1, v_2
v_3
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2$
u_1
MT
MP
$v_1 \to e_1, v_1,$ $v_2 \to e_2,$ $\neg e_2, \neg v_2$
v_2
$v_1 \to e_1, v_1,$ $v_2 \to e_2,$ $e_1, \neg e_2$
v_1
MP
MT
v_1, e_1
w
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2,$ $e_1, \neg v_2$
z_1
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2, v_2$
u_2
MP
$v_1 \to e_1, v_1,$ $v_2 \to e_2, \neg e_2,$ e_1, v_2
v_3

Figure 3.1: Left: the original model for Alice, who has not yet unpacked all consequences of her knowledge, entertaining an incomplete (u_1) and an inconsistent (u_2) world. Right: the updated model, following the applications of MP, MT.

As a result, $M, w \models \langle MP \rangle \langle MT \rangle (Ke_1 \wedge K(\neg v_2))$. In this example, we witness that the agent successfully unpacks logical consequences of her knowledge. This is because she had the necessary rules and sufficient resources to make use of them.

We can similarly model cases of reasoning failures. For example, empirical evidence suggests that MT is more cognitively costly than MP, e.g. because it is not a primitive rule (Rips, 1994) or because it requires the checking of more mental models (Johnson-Laird et al., 1992). So the cost of MT exceeds the cost of MP, and it might be the case that it is so costly for a cognitively exhausted agent to apply that she cannot do so, e.g. under time pressure (consider subjects given the abstract selection task to complete in a specific time frame). Such scenarios exemplify that the framework has tools that can fit reasoning tasks studied in psychology of reasoning. It can thus account for obstacles that may prevent agents from performing adequately in reasoning tasks which involve applications of deductive steps in a structured environment and following accumulated cognitive fatigue.

3.3 Reduction and axiomatization

3.3.1 Semantic and syntactic approaches

In this section, we reduce RSMs to possible-worlds structures augmented by syntactic functions. These functions capture the effect of impossible worlds in epistemic accessibility. The reduced structures resemble the awareness structures of Fagin and Halpern (1987). Besides, a rough division of responses to logical omniscience is between syntactic and semantic ones. It has been claimed that a syntactic approach lacks the elegance of a semantic (impossible-worlds) one, yet the latter's semantic rules do not adequately capture intuitions about knowledge and belief (Fagin and Halpern, 1987).

Notwithstanding the division, Wansing (1990) showed how various structures, like awareness-ones, can be reduced to impossible-worlds models validating precisely the same formulas (given a fixed background language). This correspondence is also displayed in the results of Thijsse (1993); Fagin et al. (1995); Halpern and Pucella (2011). Moreover, it has been extended to other variants appealing to awareness or impossible-worlds semantics. For example, Sillari (2008) showed the equivalence between logics interpreted over impossible-worlds neighborhood structures (generalizing the simple impossible-worlds structures) and over awareness neighborhood structures (generalizing the simple awareness structures), both at the propositional and the predicate level.

Our own framework falls under the semantic approaches, using impossible worlds to do justice to a non-ideal agent. Still, in order to restore explanatory power and intuitiveness, we limited the arbitrariness of impossible worlds (via Succession and Minimal Consistency) and modelled logically competent agents, in that they gradually refine their epistemic state through rule-applications. The reduction we provide (from impossible-worlds models to syntactic structures) follows the converse direction to Wansing (1990) and fits well in the picture of correspondence between the two types of attacks against logical omniscience.

Apart from showing that a richer, resource-sensitive attempt confirms this established pattern, the reduction is instrumental in providing a sound and complete logic, as it allows for the use of standard Modal Logic techniques. In this way, we wish to harvest both the benefits of this impossible-worlds semantics and the more convenient technical treatment of syntactic approaches.

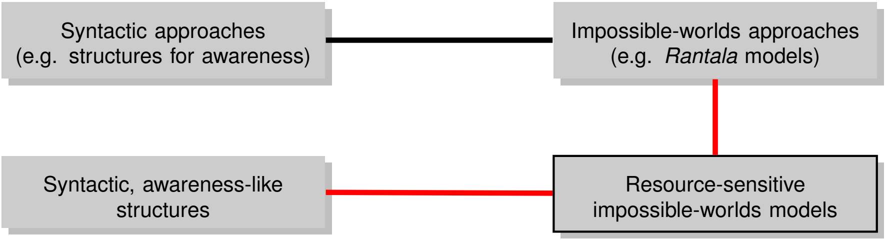

Syntactic approaches
(e.g. structures for awareness)
Impossible-worlds approaches
(e.g. *Rantala* models)
Syntactic, awareness-like structures
Resource-sensitive impossible-worlds models

Figure 3.2: The upper level depicts the correspondence between (plain) syntactic and impossible-worlds structures, as established in the literature. The bottom level and the red links depict our contribution and the analogous correspondence, which we seek to establish.

An outline of the method is as follows. First, we focus on the static part: we show that the effect of impossible worlds in the interpretation of K can be captured in a possible-worlds model, provided that suitable syntactic functions are introduced. Second, we give a sound and complete static axiomatization, through the use of well-known techniques (namely, canonical model construction). Third, we move to the dynamics and provide *reduction axioms*, that eventually reduce $\langle \rho \rangle$-involving formulas to formulas containing no such operator.

3.3.2 Reduction

Common background (static) language. We fix an appropriate language $\mathcal{L}_K^{red}$ as the "common ground" to show that the reduction is successful, i.e. that the same formulas are valid under the original and the reduced models. To that end, we introduce auxiliary operators to the static fragment of $\mathcal{L}_K$. These are such to discern the impact of possible and impossible worlds and to encode our model's structure. These operators, along with their interpretation at possible worlds, are given by:

$$M, w \models L\phi \text{ iff for all } u \in W^P \cap f(w) : M, u \models \phi$$
$$M, w \models I\phi \text{ iff for all } u \in W^I \cap f(w) : M, u \models \phi$$
$$M, w \models [RAD]_\rho \phi \text{ iff for all } u \in w^\rho\text{: } M, u \models \phi$$

The auxiliary operators L and I break down the quantifications over possible and impossible worlds involved in the interpretation of K. Operators of the form $[RAD]_\rho$ encode the model's structure as temporal-style connections generated by inference rules. This is why their interpretation should be independent of the distinction between possible and impossible worlds and thus the foregoing clause applies regardless of the world of evaluation. We also need the abbreviations below:

(a) We use $\bot$ as an abbreviation for the formula that is never true.
(b) If ϕ is of the form $\neg\psi$, for some formula ψ, then $\overline{I}\phi := I\psi$, else $\overline{I}\phi := \bot$

The reduced model. We now construct the candidate for the reduced model $\mathbf{M}_M$ given a RSM $M = \langle W^P, W^I, f, V_P, V_I, R, cp\rangle$. Take $V_I^{red}(w) := \{\phi \in \mathcal{L}_K^{red} \mid M, w \models \phi\}$, for $w \in W^I$. To capture the effect of impossible worlds in a possible-worlds structure we first need to construct a suitable *awareness-like function*:

> $\mathrm{I} : W^P \to \mathcal{P}(\mathcal{L}_K^{red})$ such that $\mathrm{I}(w) = \bigcap\limits_{v \in f(w) \cap W^I} V_I^{red}(v)$. Intuitively, I takes a possible world w and yields the set of those formulas that are true at all impossible worlds accessible from w.

3.3.1. Definition (Awareness-like structure). Given RSM $M = \langle W^P, W^I, f, V_P, V_I, R, cp\rangle$, its candidate reduced model, called *awareness-like structure* (ALS) is $\mathbf{M}_M := \langle \mathrm{W}, \mathrm{f}, \mathrm{V}, \mathrm{R}, cp, \mathrm{I}\rangle$ where:

$$\begin{aligned} &\mathrm{W} = W^P && \mathrm{f}(w) = f(w) \cap \mathrm{W} \text{ for } w \in \mathrm{W} \\ &\mathrm{V}(w) = V_P(w) \text{ for } w \in \mathrm{W} && \mathrm{R}(w) = R(w) \text{ for } w \in \mathrm{W} \\ &cp \text{ is as in the original} && \mathrm{I} \text{ is as explained before} \end{aligned}$$

The index M may be omitted if it is easily understood. Due to the construction described in Definition 3.3.1, the conditions (1-5) of epistemic RSMs translate to conditions of the corresponding ALSs. It is straightforward to see how Reflexivity, Soundness of Rules, and Rule-availability, are inherited by the ALS.

Because of Minimal Consistency, whenever $W^I \cap f(w) \neq \varnothing$ for $w \in W^P$, $\{\phi, \neg\phi\} \not\subseteq \mathrm{I}(w)$. The effect of Succession on the corresponding ALS similarly translates into conditions on the awareness-like function. More specifically, Succession affects the composition of the ρ-radius for any $\rho \in \mathcal{L}_R$. The effect of this on the corresponding ALS is reflected on the membership of $[RAD]_\rho$-formulas in the awareness-like sets $\mathrm{I}(w)$. That is:

- $\phi \in \mathrm{I}(w)$ implies $[RAD]_\rho\phi \in \mathrm{I}(w)$
- if (a) $pr(\rho) \subseteq \mathrm{I}(w)$, (b) $\neg con(\rho) \notin \mathrm{I}(w)$, and (c) $con(\rho) \neq \neg\phi$ for all $\phi \in \mathrm{I}(w)$, then $[RAD]_\rho\phi \in \mathrm{I}(w)$ implies $\phi \in \mathrm{I}(w)$, for $\phi \neq con(\rho)$
- $pr(\rho) \subseteq \mathrm{I}(w)$ implies $[RAD]_\rho con(\rho) \in \mathrm{I}(w)$
- if $pr(\rho) \subseteq \mathrm{I}(w)$ and ($\neg con(\rho) \in \mathrm{I}(w)$ or $con(\rho) = \neg\phi$ for some $\phi \in \mathrm{I}(w)$) then $[RAD]_\rho\bot \in \mathrm{I}(w)$

We name *epistemic* ALSs those structures that correspond to epistemic RSMs. The interpretation of terms in $\mathbf{M}$ is as in Definition 3.2.7, for it depends on the parameter c and on the model component cp, which are as in the original model. The clauses based on $\mathbf{M}$ are:

$$\mathbf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z \text{ iff } z_1 s_1^{\mathbf{M}} + \ldots + z_n s_n^{\mathbf{M}} \geq z$$

$\mathbf{M}, w \models p$ iff $p \in \mathrm{V}(w)$
$\mathbf{M}, w \models \neg\phi$ iff $\mathbf{M}, w \not\models \phi$
$\mathbf{M}, w \models \phi \wedge \psi$ iff $\mathbf{M}, w \models \phi$ and $\mathbf{M}, w \models \psi$
$\mathbf{M}, w \models A\rho$ iff $\rho \in \mathrm{R}(w)$

$\mathbf{M}, w \models L\phi$ iff for all $u \in \mathrm{f}(w)$: $\mathbf{M}, u \models \phi$
$\mathbf{M}, w \models I\phi$ iff $\phi \in \mathrm{I}(w)$
$\mathbf{M}, w \models [RAD]_\rho\phi$ iff $\mathbf{M}, w \models \phi$
$\mathbf{M}, w \models K\phi$ iff $\mathbf{M}, w \models L\phi$ and $\mathbf{M}, w \models I\phi$

3.3.2. Theorem (Reduction). *Given a RSM M, construct $\mathbf{M}$ as described in Definition 3.3.1. Then $\mathbf{M}$ is a reduction of M, i.e. for any $w \in W^P$ and formula $\phi \in \mathcal{L}_K^{red}$: $M, w \models \phi$ iff $\mathbf{M}, w \models \phi$.*

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

- For $\phi := p$: $M, w \models p$ iff $p \in V_P(w)$ iff $p \in \mathrm{V}(w)$ iff $\mathbf{M}, w \models p$.
- For inequalities, $\neg$, $\wedge$ and A, the claim is straightforward because the semantic clauses are the same.
- For $\phi := L\psi$: $M, w \models L\psi$ iff $M, u \models \psi$ for all $u \in W^P$ such that $u \in f(w)$ iff (by I.H.) $\mathbf{M}, u \models \psi$ for all $u \in \mathrm{W}$ such that $u \in \mathrm{f}(w)$ iff $\mathbf{M}, w \models L\psi$.
- For $\phi := I\psi$: $M, w \models I\psi$ iff $M, u \models \psi$ for all $u \in W^I$ such that $u \in f(w)$ iff $\psi \in V_I^{red}(u)$ for all $u \in W^I$ such that $u \in f(w)$ iff $\psi \in \mathrm{I}(w)$ iff $\mathbf{M}, w \models I\psi$.
- For $\phi := [RAD]_\rho\psi$: $M, w \models [RAD]_\rho\psi$ iff for all $v \in w^\rho$: $M, v \models \psi$ iff $M, w \models \psi$ iff (by I.H.) $\mathbf{M}, w \models \psi$ iff $\mathbf{M}, w \models [RAD]_\rho\psi$.
- For $\phi := K\psi$: $M, w \models K\psi$ iff $M, u \models \psi$ for all $u \in W$ such that $u \in f(w)$. Since $u \in W^P \cup W^I$, this is the case iff $M, w \models L\psi$ and $M, w \models I\psi$. Given the previous steps of the proof, this is the case iff $\mathbf{M}, w \models L\psi$ and $\mathbf{M}, w \models I\psi$, iff $\mathbf{M}, w \models K\psi$.

□

3.3.3 Static axiomatization

We first present an axiomatic system for the static part and show that it is sound and complete with respect to the reduced models.

3.3.3. Definition (Axiomatization of Λ_K). The static logic Λ_K is axiomatized by Table 3.1 and the rules *Modus Ponens* and *Necessitation$_L$* (from ϕ, infer $L\phi$).

INEQ, appearing in Fagin et al. (1990); Fagin and Halpern (1994); Halpern (2017), is introduced to accommodate the inequalities. The axiom K_L reflects that the auxiliary operator L behaves as K does in standard modal-epistemic logics. Similarly, T_L captures the factivity of knowledge and corresponds to the Reflexivity of the epistemic models. The axioms SoR, MC, RA correspond to our respective model conditions (Minimal Consistency, Soundness of rules, and Rule-availability), given how these are reflected on our language. The axioms under

PC	All instances of classical propositional tautologies
INEQ	All instances of valid formulas about linear inequalities
K_L	$L(\phi \to \psi) \to (L\phi \to L\psi)$
T_L	$L\phi \to \phi$
SoR	$A\rho \to tr(\rho)$
MC	$I\bot \lor (\neg(I\phi \land I\neg\phi))$
RA	$A\rho \to LA\rho$
SUCC	$I\phi \to I[RAD]_\rho\phi$
	$\bigwedge\limits_{\psi \in pr(\rho)} I\psi \land \neg I\neg con(\rho) \land \neg\overline{I}con(\rho) \to (I[RAD]_\rho\phi \to I\phi),\ \phi \neq con(\rho)$
	$\bigwedge\limits_{\psi \in pr(\rho)} I\psi \to I[RAD]_\rho con(\rho)$
	$\bigwedge\limits_{\psi \in pr(\rho)} I\psi \land (I\neg con(\rho) \lor \overline{I}con(\rho)) \to I[RAD]_\rho\bot$
RAD	$[RAD]_\rho\phi \leftrightarrow \phi$
RED	$K\phi \leftrightarrow L\phi \land I\phi$

Table 3.1: The static axioms

SUCC correspond to the effect of Succession, as reflected on the behaviour of ρ-expansions of impossible worlds. The axiom RAD corresponds to the behaviour of ρ-expansions of possible worlds while RED reduces K in terms of the auxiliary operators L and I.

3.3.4. Theorem (Soundness). *The logic Λ_K is sound with respect to epistemic ALSs.*

Proof:
It suffices to show that the axioms are valid in this class. The claims for PC, INEQ are straightforward, as is for K_L. The axioms for T_L, SoR, MC, RA, SUCC, and RAD are valid due to the model conditions. The validity for RED follows from the constructions of the **M**-semantic clauses for L and I. □

3.3.5. Theorem (Completeness). *The logic Λ_K is complete with respect to epistemic ALSs.*

Proof:
We have to show that every Λ_K-consistent set is satisfiable in a structure of the given class. Following (Blackburn et al., 2001, Chapter 4), we aim at constructing a suitable canonical model corresponding to our class of structures (epistemic ALSs). Notice that taking (maximal) Λ_K-consistent sets and showing Lindenbaum's lemma (*If Γ is a Λ_K-consistent set of formulas then there is a maximal*

Λ_K*-consistent set* Γ^+ *such that* $\Gamma \subseteq \Gamma^+$) go as usual in the literature. Then, our canonical model $\mathcal{M}$ consists of:

$\mathcal{W}$ the set of all maximal Λ_K-consistent sets

$\mathcal{F}$ such that for $w, u \in \mathcal{W}$: $u \in \mathcal{F}(w)$ iff $\{\phi \mid L\phi \in w\} \subseteq u$

$\mathcal{V}(w) = \{p \mid p \in w\}$

$\mathcal{R}(w) = \{\rho \mid A\rho \in w\}$

cp as before

$\mathcal{I}(w) = \{\phi \mid I\phi \in w\}$

The axioms indicating model properties ensure that the canonical model is in the class of epistemic ALSs. More specifically, T_L ensures Reflexivity, SoR Soundness of Rules, MC Minimal Consistency, RA Rule-availability, SUCC Succession, and finally RAD ensures that the radius of a possible world is itself.

It then suffices to show the truth lemma (i.e. $\mathcal{M}, w \models \psi$ iff $\psi \in w$). We do so by induction on ψ.

- For $\psi := p$: $\mathcal{M}, w \models p$ iff $p \in \mathcal{V}(w)$ iff $p \in w$ by definition of $\mathcal{V}$.
- The claims for Boolean connectives, linear inequalities, $[RAD]_\rho$, and A follow directly from the I.H., INEQ, RAD, and the construction of the canonical model (namely, properties of maximal consistent sets and $\mathcal{R}$).
- For $\psi := L\phi$: $\mathcal{M}, w \models L\phi$ iff $\mathcal{M}, u \models \phi$ for all $u \in \mathcal{W}$ such that $u \in \mathcal{F}(w)$ iff (by I.H) $\phi \in u$ for all $u \in \mathcal{W}$ such that $u \in \mathcal{F}(w)$.

  As a result, we have to show that $L\phi \in w$ iff $\phi \in u$ for all $u \in \mathcal{W}$ such that $u \in \mathcal{F}(w)$.

  ▷ The left-to-right direction follows from the definition of $\mathcal{F}$.

  ▷ For the other direction, we suppose that $L\phi \notin w$ and we have to show that there is some $u \in \mathcal{W}$ such that $u \in \mathcal{F}(w)$ and $\phi \notin u$. By the definition of $\mathcal{F}$ and maximal consistency, this amounts to: there is some $u \in \mathcal{W}$ such that $\{\chi \mid L\chi \in w\} \subseteq u$ and $\neg\phi \in u$, i.e. $\{\chi \mid L\chi \in w\} \cup \{\neg\phi\} \subseteq u$. It suffices to show that $\{\chi \mid L\chi \in w\} \cup \{\neg\phi\}$ is Λ_K-consistent. Suppose it is not. Then $\vdash \chi_1 \wedge \ldots \wedge \chi_n \rightarrow \phi$ for some $\{L\chi_1, \ldots, L\chi_n\} \subseteq w$. But due to maximal consistency $(L\chi_1 \wedge \ldots \wedge L\chi_n \rightarrow L\phi) \in w$ and since $L\chi_1 \wedge \ldots \wedge L\chi_n \in w$, $L\phi \in w$ too. This contradicts the initial hypothesis. Therefore, the set is Λ_K-consistent, as desired.
- For $\psi := I\phi$: $\mathcal{M}, w \models I\phi$ iff $\phi \in \mathcal{I}(w)$ iff $I\phi \in w$ by construction of $\mathcal{I}$.
- For $\psi := K\phi$: $\mathcal{M}, w \models K\phi$ iff $\mathcal{M}, w \models L\phi \wedge I\phi$ iff $\mathcal{M}, w \models L\phi$ and $\mathcal{M}, w \models I\phi$ iff (by I.H.) $L\phi \in w$ and $I\phi \in w$ iff $L\phi \wedge I\phi \in w$ iff $K\phi \in w$.

□

3.3.4 Dynamic axiomatization

Moving to the dynamic part, we look into the behaviour of rule-applications under ALSs. Formulas of the form $\langle\rho\rangle\phi$ are interpreted as indicated by the original clause (Definition 3.2.8), only now using the ALSs corresponding to M and M^{ρ}.

In order to obtain the full axiomatization, we will follow the common DEL practice of providing reduction axioms for the dynamic operators, in our case, for $\langle\rho\rangle$. Because of this, we can eventually reduce formulas involving these operators to purely static formulas. Once dynamic formulas are translated to provably equivalent ones without the dynamic operators, we can use the completeness of the static "base" logic (Section 3.3.3) to prove completeness for the full language (van Ditmarsch et al., 2007, Chapter 7). We follow this procedure and try to reduce formulas with dynamic operators to formulas involving no such operators.

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

3.3.6. PROPOSITION ($\langle\rho\rangle$-reduction axioms). *The following formulas are valid:*

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

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

Proof:
It is easy to check that the claim follows for the atoms, the Boolean cases, A, $[RAD]_{\sigma}$, and L. The claim for inequalities follows because the abbreviation for the updated terms captures the changes of the value of capacity in the language (terms for costs are unchanged). In what follows, we show the claim for the I operator. Given the reduction axioms for L and I, that for K follows easily because the clause for K is given in terms of the auxiliary operators L and I.

It suffices to show that the reduction axiom is valid in an arbitrary epistemic RSM. Let M be such a model and w an arbitrary possible world of the model.

- Suppose $M, w \models \langle\rho\rangle I\phi$. Therefore, $M, w \models (cp \geq c_{\rho})$, $M, w \models A\rho$ and $M^{\rho}, w \models I\phi$. As a result:

  $$\text{for all } v \in f^{\rho}(w) \cap W^{I} : M^{\rho}, v \models \phi \qquad (1)$$

  Take arbitrary $u \in W^{I}$ and arbitrary $v \in u^{\rho}$. Then, $v \in f^{\rho}(w) \cap W^{I}$, and by (1) and the definitions of V_I and radius: $M, v \models \phi$. Overall, $M, w \models I[RAD]_{\rho}\phi$ and by $M, w \models (cp \geq c_{\rho})$, $M, w \models A\rho$, we finally get $M, w \models (cp \geq c_{\rho}) \wedge A\rho \wedge I[RAD]_{\rho}\phi$.

- For the other direction, suppose that $M, w \models (cp \geq c_\rho) \wedge A\rho \wedge I[RAD]_\rho\phi$. Take arbitrary $v \in f^\rho(w) \cap W^I$, i.e. there is $u \in f(w) \cap W^I$ such that $v \in u^\rho$. By the truth conditions of $M, w \models I[RAD]_\rho\phi$, we obtain $M, v \models \phi$, and by definitions of V_I and radius, $M^\rho, v \models \phi$. Overall, $M^\rho, w \models I\phi$, and finally $M, w \models \langle\rho\rangle I\phi$.

□

3.3.7. Theorem (Full axiomatization). *The logic given by the system of Definition 3.3.3, the reduction axioms of Proposition 3.3.6, with the additional* $[\rho]$*-Necessitation Rule, is sound and complete with respect to epistemic ALSs.*

Proof:
The result follows from Proposition 3.3.6, Theorem 3.3.4, Theorem 3.3.5. □

3.4 Conclusions

The proposed dynamic logical framework overcomes logical omniscience, using resource-sensitive models and suitable model updates to account for a single agent's stepwise deductive reasoning. As evinced by the formal results, these processes are bounded, dictated by empirical facts on the difficulty of deductive reasoning. Meanwhile, we argue for the adequacy of this semantic approach against logical omniscience in terms of explanatory power, but we also show that it can be reduced to a syntactic one. In doing so, we enrich the picture of the correspondence between syntactic and semantic approaches against logical omniscience and we offer a useful detour that allows for the use of standard results to extract a sound and complete logic.

Overall, this proposal contributes towards one goal we set in Chapter 2: to build logical systems motivated by the alternative picture, explicitly representing the *deductive* reasoning steps of human agents and the effort they require. In the following chapters, we elaborate on this development. In Chapter 4, we show that this resource-sensitive view can be adapted to a logical framework that goes beyond the coarse-grained mental classification criticized by Goldman (1978). In Chapter 5, we show that these results can contribute to the logical modelling of the dual process theories of reasoning.

Chapter 4

Plausibility models and bounded agents

Plausibility models are variants of the relational Kripke models and they have been used in (D)EL to account for a greater variety of propositional attitudes and policies of incorporating information. In this chapter, we revisit this type of modelling from a resource-sensitive perspective, by extending our previous results.[1]

In particular, we provide resource-sensitive variants of more fine-grained attitudes through the use of suitable plausibility models. These models are supplemented by (a) impossible worlds, 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. Deductive reasoning is reflected on the dynamic truth clauses. These include resource-sensitive "preconditions" and utilize a model update mechanism that modifies the set of accessible worlds and their plausibility, but also reduces cognitive capacity by the appropriate cost. We further show that our plausibility models can be reduced to awareness-like plausibility structures that validate the same formulas, and we then 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. This allows us to discuss tasks that combine bounded reasoning with the revision of epistemic and doxastic states that occurs when the agent hears or observes new external information.

The chapter is structured as follows. In Section 4.1, we summarize why and how plausibility models have been used in the DEL literature. We then introduce our framework and discuss its contribution to the highlighted topics in Section 4.2. The reduction procedure and the axiomatization are given in Section 4.3. In Section 4.4, we explain how the framework is combined with the dynamics of interaction, and we finally conclude in Section 4.5.

[1] The chapter is based on Smets and Solaki (2018); Solaki and Smets (2021); Solaki (2019).

4.1 Background

The first logical systems within DEL, which were designed to model epistemic updates, were followed by more sophisticated theories that have been developed to represent a variety of informational changes including epistemic updates, doxastic changes, preference change, etc. Within this variety of systems, the tools that we need to represent an agent's mental states are called plausibility models (Grove, 1988; Board, 2004; van Benthem, 2007; Baltag and Smets, 2008b). Plausibility models allow the study of nuanced epistemic and doxastic attitudes and facilitate the introduction of a repertoire of epistemic and doxastic actions.

4.1.1. Definition (Plausibility model). A *plausibility model* M is a structure $\langle W, \geq, V \rangle$ where:

- W is a non-empty set of possible worlds.
- $\geq$ is a *locally well-preordered (plausibility) relation* on W, such that $w \geq u$ reads "w is considered no more plausible than u".
- V is a valuation such that each world is assigned to a set of propositional atoms from a given set Φ (those true at the world).

Between any two worlds entertained by the agent as ways things could be, there is a (relative) plausibility ordering. The ordering is a local well-preordering, which means that $\geq$ is reflexive, transitive, locally connected (i.e. the relation is connected in each comparability class), and converse wellfounded (i.e. there is no infinite ascending $\geq$-chain thus a set of *most* plausible worlds can always be retrieved) (Baltag and Renne, 2016). A pair (M, w) consisting of a model M and a designated world w of the model (taken as the actual world from the perspective of the modeller), is called a *pointed plausibility model*.

Plausibility models allow us to characterize a variety of epistemic and doxastic attitudes, which includes, besides the strong concept of knowledge used in static EL (i.e. knowledge as truth in all accessible possible worlds), weaker epistemic attitudes as well (Baltag and Smets, 2011, 2013). In Baltag and Smets (2008b), one such weaker attitude is coined "safe belief" or "(in)defeasible knowledge" referring to the epistemic concept described by Lehrer and Paxson (1969); Lehrer (2000); Stalnaker (2006). If we explain this notion of defeasible knowledge in terms of the extra ingredients one needs to add to a notion of belief, the most straightforward way is to refer to a *stability*-account (Rott, 2004). Using "stability" as the identifying factor, defeasible knowledge refers to an agent's justified true belief that is stable when any new true information is received.[2] We will

[2] If, as Floridi (2005) claims, information is factive, there cannot be "false" information. Works on belief revision, however, generally adopt a weaker information sense, whereby (declarative) information is taken to be meaningful data, not perforce truthful (van Benthem, 2011).

follow this DEL literature and represent defeasible knowledge by a modal operator $\Box$ (as in Baltag and Smets (2008b)). The truth conditions for $\Box\phi$, when evaluated at a world in a plausibility model, ask for ϕ to hold at all worlds that are at least as plausible as the point of evaluation. The truth conditions for $B\phi$ require that ϕ holds at the set of most plausible worlds of the model, denoted by $min(W)$.[3] While K represents an agent's fully introspective and factive attitude, $\Box$ is factive but not fully introspective. This weaker notion usually satisfies the S4-properties, while K satisfies the S5-properties and is considered to be infallible and indefeasible.

This more graded outlook of different attitudes fits well with the distinctions of attitudes advocated as desirable ingredients of an alternative rationality picture in Section 2.4.3. Besides, our attitude towards a piece of information is oftentimes not as strong as the strong concept of infallible knowledge nor as weak as plain belief. Moving forward, we show how resource-sensitive modelling is fully compatible with the fine-grained understanding of propositional attitudes and their changes.

4.2 Plausibility and resource-boundedness

4.2.1 Syntax

The syntax for this enriched framework builds on the definitions of inference rules (Definition 3.2.1) and terms (Definition 3.2.2). We introduce the language $\mathcal{L}_{K,\Box}$, extending $\mathcal{L}_K$ (Definition 3.2.3) with an additional epistemic modality. Apart from K for conventional knowledge, we use $\Box$ to express defeasible knowledge. Regarding the changes of the agent's epistemic state induced by deductive reasoning, we use dynamic operators labelled by inference rules (of the form $\langle\rho\rangle$) as before.

4.2.1. Definition (Language $\mathcal{L}_{K,\Box}$). The language $\mathcal{L}_{K,\Box}$ is defined by:

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

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

4.2.2 Resource-sensitive plausibility models

Our semantics is based on a special type of plausibility models. In line with Definition 3.2.4, the model is augmented by impossible worlds, restrained by a requirement of Minimal Consistency. While the agent's fallibility is not precluded

[3] One can further define *conditional belief* in terms of the two forms of knowledge we discussed, i.e. both the strong and weaker notion (van Ditmarsch et al., 2015; Baltag and Renne, 2016). This is instrumental in capturing so-called *static belief change*, as it expresses what we believe *conditional* to some other piece of information.

– 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. To capture which worlds are considered plausible by the agent and how so, we use a mapping from a given set of worlds to the class of ordinals Ω to derive the plausibility ordering (in line with Spohn (1988)). The models are parameterized by the set of resources *Res* and the cognitive cost function c, as described in Section 3.2.2. Rule-availability and cognitive capacity, which is decreased appropriately by the cost of each reasoning step, are also components of the models as before.

4.2.2. Definition (Resource-sensitive plausibility model). Given a set of r-many resources *Res* and a cost function c, a *resource-sensitive plausibility model* (RSPM) is a tuple $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp \rangle$ where:

- W^P, W^I are non-empty sets of possible and impossible worlds respectively.
- $W := \{w \mid w \in W^P \text{ or } w \in W' \subseteq W^I\}$ is the set of worlds entertained by the agent.
- *ord* is a function from W to the class of ordinals Ω assigning an ordinal to each world.
- $V_P : W^P \to \mathcal{P}(\Phi)$ is a valuation function assigning to each *possible* world, the propositional *atoms* that are true there.
- $V_I : W^I \to \mathcal{P}(\mathcal{L}_{K,\Box})$ is a valuation function assigning to each *impossible* world, the formulas (*atomic or complex*) that are true there.
- $R : W^P \to \mathcal{P}(\mathcal{L}_R)$ is a function yielding the rules the agent has available (i.e. has acknowledged as truth-preserving) at each possible world.
- *cp* denotes the agent's cognitive capacity, i.e. $cp \in \mathbb{Z}^r$, intuitively standing for what the agent can afford with respect to each resource.

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, (locally) connected and conversely well-founded.[4] 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 (van Ditmarsch et al., 2015, Chapter 7). For reasons described in Section 3.2.2, we impose:

(1) **Minimal Consistency**: $\{\phi, \neg\phi\} \not\subseteq V_I(w)$ for all $w \in W^I, \phi \in \mathcal{L}_{K,\Box}$

[4] 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.

To ensure that the rules available to the agent are truth-preserving, and assuming that propositional formulas are evaluated as usual at possible worlds, we impose:

(2) **Soundness of rules**: for $w \in W^P$, if $\rho \in R(w)$ then $M, w \models tr(\rho)$, where
$$tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \to con(\rho)$$

As in Section 3.2.2, we ask:

(3) **Rule-availability**: for $w, u \in W^P$, if $\rho \in R(w)$ then $\rho \in R(u)$

We also need a condition to hardwire the effect of deductive reasoning in the model. To that end, we ask:

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

We call u a *ρ-expansion* of w. For $w \in W^P$, its *ρ-expansion* is w itself, due to deductive closure of possible worlds. Based on this, we define the ρ-radius of a world exactly as in Definition 3.2.5. Henceforth, whenever we refer to RSPMs, we refer to models fulfilling the aforementioned conditions.

4.2.3 Model transformations and truth clauses

To evaluate $\langle\rho\rangle\phi$, we have to examine the truth value of ϕ in a transformed model, defined in a way that captures the effect of applying ρ. Roughly, a pointed model (M', w) is the *ρ-update* of a given pointed RSPM (M, w), whenever the set of worlds entertained by the agent 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 $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$ be a RSPM. Given a rule ρ, the updated M^ρ is given by $\langle W^P, W^I, W^\rho, ord^\rho, V_P, V_I, R, cp^\rho\rangle$ designed as follows:

Step 1 The set of entertained worlds is $W^\rho := \bigcup_{v \in W} v^\rho$. In words, W^ρ consists of 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 impossible worlds entertained by the agent.

Step 2 We now develop the new ordering ord^ρ 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 $u \in v^\rho$. 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. Again, for $u, v \in W^\rho$, we say: $u \geq^\rho v$ iff $ord^\rho(u) \geq ord^\rho(v)$.

Step 3 $cp^\rho := cp - c(\rho)$.

It is easy to check that all the conditions on RSPMs are preserved. The terms are interpreted as in Definition 3.2.7. Then the plausibility clauses are:

4.2.3. Definition (Plausibility truth clauses). The following clauses inductively define when a formula ϕ is *true at w in M* (notation: $M, w \models \phi$). For $w \in W^I$: $M, w \models \phi$ iff $\phi \in V_I(w)$. For $w \in W^P$, given that the Boolean cases are standard:

$$\begin{aligned}
&M, w \models p \text{ iff } p \in V_P(w)\\
&M, w \models z_1 s_1 + \ldots + z_n s_n \geq z \text{ iff } z_1 s_1^M + \ldots + z_n s_n^M \geq z\\
&M, w \models K\phi \text{ iff } M, u \models \phi \text{ for all } u \in W\\
&M, w \models \Box\phi \text{ iff } M, u \models \phi \text{ for all } u \in W \text{ such that } w \geq u\\
&M, w \models A\rho \text{ iff } \rho \in R(w)\\
&M, w \models \langle\rho\rangle\phi \text{ iff } M, w \models cp \geq c_\rho,\ M, w \models A\rho \text{ and } M^\rho, w \models \phi
\end{aligned}$$

Validity is again defined with respect to possible worlds.

4.2.4 Discussion

With these constructions, we overcome logical omniscience while still accounting for how agents perform inferences lying within suitable applications of rules. Importantly, these results are generalized to a greater spectrum of attitudes and can be further adapted to provide resource-sensitive variants of other notions (strong belief, conditional belief, etc.) that can be captured via plausibility. This complies with the sub-goal put forward in Section 2.5.3 and addresses the aspects Goldman (1978) delineates for the fruitful connection of epistemology and psychology. In particular, the argument of impossible worlds suffices to invalidate the closure principles. The truth conditions for $\langle\ddagger\rangle\spadesuit\phi$, where $\langle\ddagger\rangle$ abbreviates a sequence of rules and $\spadesuit$ stands for a propositional attitude such as K or $\Box$, demonstrate that an agent can come to refine her state by following an *affordable* and *available* reasoning track. As before, the rule-sensitivity, the measure on cognitive capacity, and the way it is updated allow us to determine the evolution of a reasoning track. The extent of the agent's "feasible inferences" is informed by a mechanism that imports descriptive facts on the difficulty of different reasoning steps.

4.2.4. Example (Bounded muddy children). 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 of 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.

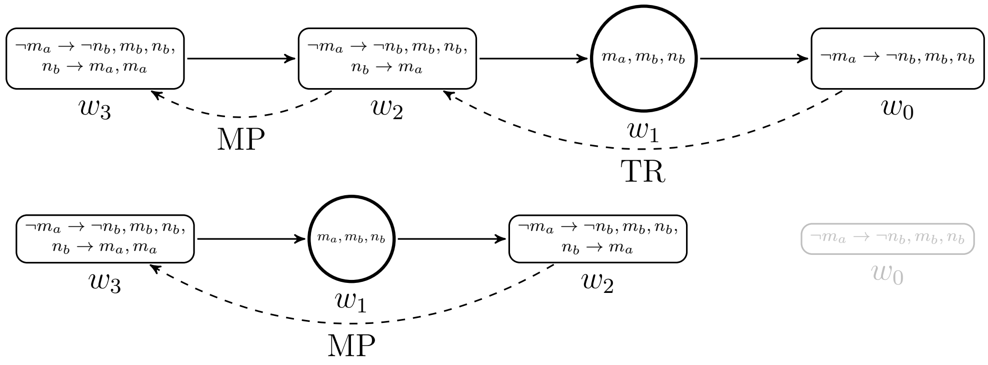
In this example, we analyze the failure of applying a sequence of rules in the $k = 2$ scenario, attributed to the fact that the first rule applied is so cognitively costly for a child that her available time expires before she can apply the next. It thus becomes clear why in even more complex cases (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 can model the dynamics of inference and the resource consumption each step induces.

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". For simplicity, take two rules, Transposition of the implication (TR) and Modus Ponens (MP), such that $TR = \{\neg m_a \to \neg n_b\} \rightsquigarrow n_b \to m_a$, $MP = \{n_b, n_b \to m_a\} \rightsquigarrow m_a$, $Res = \{time, memory\}$, $c(TR) = (5,2), c(MP) = (1,2)$, $cp = (5,7)$. Let $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$ be as depicted in Figure 4.1 (top). That is, $W^P = \{w_1\}$, $W^I = \{w_3, w_2, w_0\}$, $W = W^P \cup W^I$, ord is understood by the index of each world, and $R(w_1) = \{TR, MP\}$.[5]

Analyzing the reasoning of child a, 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 that b said "no". Following a TR-application, the world w_0 is eliminated from W and its position is taken by its TR-expansion (w_2) and $cp^{TR} = (5,7) - (5,2) = (0,5)$ (Figure 4.1

[5] The same conventions apply as in Example 3.2.9. Circles are used for deductively closed possible worlds (W^P) and rectangles for impossible worlds (W^I), where we write all formulas satisfied. We indicate (non-trivial) expansions via dashed arrows and the plausibility ordering via plain arrows. The reflexive and transitive arrows have been omitted for simplicity.

(bottom)). 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 valid.



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

Figure 4.1: The reasoning of boundedly rational child a from M (top) to M^{TR} (bottom).

4.3 Reduction and axiomatization

In this section, we extend the reduction and axiomatization results of Section 3.3 to the plausibility framework. We have seen that the resource-sensitive, impossible-worlds semantics can be brought together with plausibility modelling. We will now show that this framework, although richer in that it accommodates finer attitudes, still fits in the correspondence picture between semantic and syntactic approaches. More specifically, RSPMs can be reduced to syntactic plausibility structures. In the absence of impossible worlds, we can use standard techniques used in axiomatizing DEL settings. This is a technical contribution; the components of the reduced model may lack the intuitive readings of the original one (e.g. impossible worlds as witnesses of the agent's fallibility), yet they allow us to prove soundness and completeness. Therefore, this method has the advantage of combining the benefits of impossible worlds in plausibility models and the technical treatment facilitated by awareness-like 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 operators 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 without them. Because of that, we can appeal to the completeness of the static part to get a complete axiomatization for the full setting.

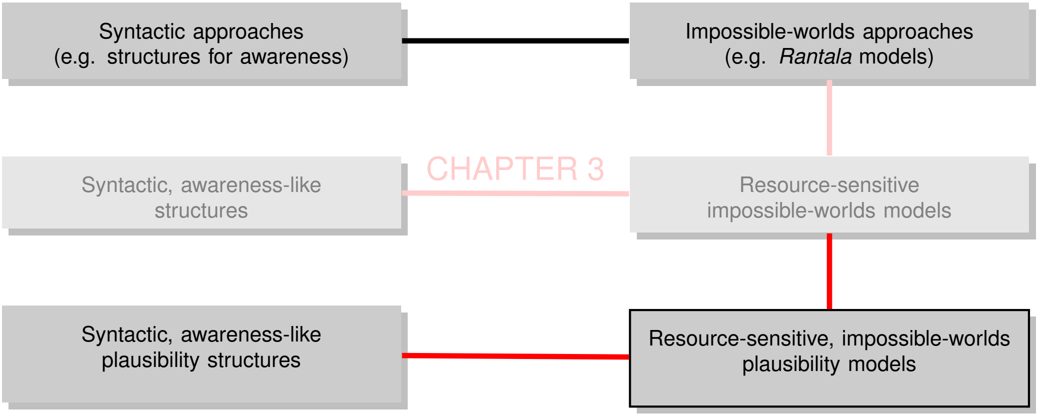

Syntactic approaches
(e.g. structures for awareness)
Impossible-worlds approaches
(e.g. *Rantala* models)
CHAPTER 3
Syntactic, awareness-like
structures
Resource-sensitive
impossible-worlds models
Syntactic, awareness-like
plausibility structures
Resource-sensitive, impossible-worlds
plausibility models

Figure 4.2: Extending the reduction to the new plausibility framework

4.3.1 Reduction

The static background language. We first need to fix the "common ground", an appropriate language $\mathcal{L}^{red}_{K,\Box}$ to show that the same formulas are valid under the original RSPMs and the reduced models. Since the methodology for breaking down K and $\Box$ into auxiliary operators will be the same, we will use $\spadesuit$ as an abbreviation standing for K or $\Box$. Then, take the *quantification set* $Q_\spadesuit(w)$ to be W if $\spadesuit = K$, and $\{u \in W \mid w \geq u\}$, 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 $\mathcal{L}_{K,\Box}$, in order to capture (syntactically) the effect of impossible worlds in the interpretations of propositional attitudes. For $w \in W^P$:

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

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:

$$M, w \models [RAD]_\rho \phi \text{ iff for all } u \in w^\rho\text{: } M, u \models \phi$$

The operators $[RAD]_\rho$, labelled by inference rules are such to ensure that all ρ-expansions are ϕ-satisfying. Indexed operators of this form provide information on the model's structure and their interpretation is independent of the distinction between possible and impossible worlds. These auxiliary operators essentially function like those of Section 3.3.2, only now in the context of plausibility models. In the same spirit, we also use the following abbreviation: if ϕ is of the form $\neg\psi$, for some formula ψ, then $\overline{I}_\spadesuit \phi := I_\spadesuit \psi$, else $\overline{I}_\spadesuit \phi := \bot$.

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

- $\mathrm{I}_\spadesuit : W^P \to \mathcal{P}(\mathcal{L}_{K,\Box}^{red})$ such that $\mathrm{I}_\spadesuit(w) = \bigcap_{v \in W^I \cap Q_\spadesuit(w)} V_I^{red+}(v)$. 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.

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 later developed.

4.3.1. Definition (Awareness-like plausibility structure). Given an original RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$, our candidate reduced model, called *awareness-like plausibility structure* (ALPS), is the tuple $\mathbf{M}_M = \langle \mathrm{W}, \geq, \sim, \mathrm{V}, \mathrm{R}, cp, \{\mathrm{I}_\spadesuit\}_{\{\spadesuit = K,\Box\}}\rangle$ where:

$\mathrm{W} = W^P$	$\mathrm{V}(w) = V(w)$ for $w \in \mathrm{W}$
$u \geq w$ iff $ord(u) \geq ord(w)$, for $w, u \in \mathrm{W}$	$\mathrm{R}(w) = R(w)$ for $w \in \mathrm{W}$
$u \sim w$ iff $u \geq w$ or $w \geq u$, for $w, u \in \mathrm{W}$	$\mathrm{I}_\spadesuit$ as explained before

The index M may be omitted when no confusion arises. Due to the construction of awareness-like functions, the properties of the original models are translated into properties of the reduced models. 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 interpretation of terms under an ALPS $\mathbf{M}$ is as before (Definition 3.2.7). The semantic clauses, based on $\mathbf{M}$, are standard for the Boolean connectives; the remaining are given below:

$$\mathbf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z \text{ iff } z_1 s_1^{\mathbf{M}} + \ldots + z_n s_n^{\mathbf{M}} \geq z$$

$\mathbf{M}, w \models p$ iff $p \in \mathrm{V}(w)$	$\mathbf{M}, w \models L_\spadesuit \phi$ iff $\mathbf{M}, u \models \phi$ for all $u \in \mathrm{Q}_\spadesuit(w)$
$\mathbf{M}, w \models I_\spadesuit \phi$ iff $\phi \in \mathrm{I}_\spadesuit(w)$	$\mathbf{M}, w \models [RAD]_\rho \phi$ iff $\mathbf{M}, w \models \phi$
$\mathbf{M}, w \models A\rho$ iff $\rho \in \mathrm{R}(w)$	$\mathbf{M}, w \models \spadesuit \phi$ iff $\mathbf{M}, w \models L_\spadesuit \phi$ and $\mathbf{M}, w \models I_\spadesuit \phi$

4.3.2. Theorem (Reduction). *Given a RSPM M, let $\mathbf{M}$ be the ALPS obtained as in Definition 4.3.1. Then $\mathbf{M}$ is a reduction of M, i.e. for any $w \in W^P$ and formula $\phi \in \mathcal{L}_{K,\Box}^{red}$: $M, w \models \phi$ iff $\mathbf{M}, w \models \phi$.*

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

- For inequalities, $\neg$, $\wedge$, $[RAD]_\rho$, 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 $u \in W^I \cap Q_\spadesuit(w)$: $\psi \in V_I^{red+}(u)$ 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 $\mathbf{M}, w \models \spadesuit\psi$.

□

4.3.2 Static axiomatization

We have reduced RSPMs to ALPSs. We now develop the (static) logic $\Lambda_{K,\Box}$, showing that it is sound and complete with respect to them.

4.3.3. Definition (Axiomatization of $\Lambda_{K,\Box}$). $\Lambda_{K,\Box}$ is axiomatized by Table 4.1 and the rules *Modus Ponens*, $\mathit{Necessitation}_{L_K}$ (from ϕ, infer $L_K\phi$) and $\mathit{Necessitation}_{L_\Box}$ (from ϕ, infer $L_\Box\phi$).

PC	All instances of classical propositional tautologies
INEQ	All instances of valid formulas about linear inequalities
L$_K$	The **S5** axioms for L_K
L$_\Box$	The **S4** axioms for $L_\Box$
SoR	$A\rho \to tr(\rho)$
MC	$I_\spadesuit\bot \vee (\neg(I_\spadesuit\phi \wedge I_\spadesuit\neg\phi))$
RA	$A\rho \to L_K A\rho$
LC	$L_K(\phi \vee L_\Box\psi) \wedge L_K(\psi \vee L_\Box\phi) \to (L_K\phi \vee L_K\psi)$
INDEF	$L_K\phi \to L_\Box\phi$
	$I_K\phi \to I_\Box\phi$
SUCC	$I_\spadesuit\phi \to I_\spadesuit[RAD]_\rho\phi$
	$\bigwedge_{\psi \in pr(\rho)} I_\spadesuit\psi \wedge \neg I_\spadesuit\neg con(\rho) \wedge \neg\overline{I}_\spadesuit con(\rho) \to (I_\spadesuit[RAD]_\rho\phi \to I_\spadesuit\phi)$, $\phi \neq con(\rho)$
	$\bigwedge_{\psi \in pr(\rho)} I_\spadesuit\psi \to I_\spadesuit[RAD]_\rho con(\rho)$
	$\bigwedge_{\psi \in pr(\rho)} I_\spadesuit\psi \wedge (I_\spadesuit\neg con(\rho) \vee \overline{I}_\spadesuit con(\rho)) \to I_\spadesuit[RAD]_\rho\bot$
RAD	$[RAD]_\rho\phi \leftrightarrow \phi$
RED	$\spadesuit\phi \leftrightarrow L_\spadesuit\phi \wedge I_\spadesuit\phi$

Table 4.1: The static axioms

Recall that $\spadesuit$ is an abbreviation which can be substituted for K or $\Box$. INEQ is as before (Section 3.3.3) introduced to accommodate the linear inequalities. 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 SoR, MC, RA, and SUCC take care of the respective model conditions (Soundness of Rules, Minimal Consistency, Rule-availability, and Succession) to the extent that these are reflected on the language. The same holds for INDEF and LC, which mimic the axioms for indefeasibility and local connectedness in the usual plausibility structures (Baltag and Smets, 2008b). To capture the behaviour of radius, we also introduce the RAD axiom. Finally, RED expresses K and $\Box$ in terms of the corresponding auxiliary operators. We now move to the theorems for soundness and completeness with respect to the ALPSs corresponding to our RSPMs.

4.3.4. Theorem (Soundness). *$\Lambda_{K,\Box}$ is sound with respect to ALPSs.*

Proof:
Standard arguments suffice regarding PC, INEQ, L_K, $\mathsf{L}_\Box$ and *Modus Ponens*, $Necessitation_{L_K}$, $Necessitation_{L_\Box}$ preserve validity as usual. The axioms for SoR, MC, RA, SUCC are valid due to the respective conditions placed on the model. The validity of LC is due to the connectedness of the model. The validity of INDEF and RED is a direct consequence of the semantic clauses for $\spadesuit, L_\spadesuit, I_\spadesuit$. The validity of RAD is straightforward. □

4.3.5. Theorem (Completeness). *$\Lambda_{K,\Box}$ is complete with respect to ALPSs.*

Proof:
The proof is a variant of the one in (Baltag and Smets, 2008b, p.36). Towards showing completeness, we use a suitable canonical plausibility model. Taking (maximal) $\Lambda_{K,\Box}$-consistent sets and showing Lindenbaum's lemma follow the standard procedure. The canonical model for the logic $\Lambda_{K,\Box}$ is built similarly to that one of Theorem 3.3.5, now containing:

- $\mathcal{W}$, the set of all maximal $\Lambda_{K,\Box}$-consistent sets.
- $\geq$, such that for $w, u \in \mathcal{W}$: $w{\geq}u$ iff $\{\phi \mid L_\Box\phi \in w\} \subseteq u$.
- $\sim$, 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 $w \in \mathcal{W}$.

Due to $\mathsf{L}_\Box$, LC, and Modal Logic results on correspondence (Blackburn et al. (2001)) the canonical model is reflexive, transitive and (locally) connected (with respect to $\geq$) and due to L_K and INDEF, $\sim$ is the symmetric extension of $\geq$

(these properties yield the so-called *non-standard* plausibility models). The axioms on SoR, MC, RA, SUCC, and RAD are such to ensure that the model has the corresponding special properties.

We then perform induction on the complexity of ϕ 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 $\mathcal{R}$), the I.H., INEQ, and the properties of maximal consistent sets. The claim for $[RAD]_\rho$ follows by the I.H. and RAD. The claims for $L_\Box$ and L_K follow with the help of I.H. (as in proof Theorem 3.3.5 for L) while for $I_\Box$, I_K, we rely on the construction of the awareness-like functions and then the result is immediate. For $K\psi$ and $\Box\psi$, we make use of RED, the I.H., and the results of the previous steps on L_K, I_K and $L_\Box$, $I_\Box$.

This yields completeness with respect to non-standard ALPSs. This is analogous to completeness with respect to non-standard plausibility models (Baltag and Smets, 2008b) and non-standard plausibility-access models (Velázquez-Quesada, 2014). Such models have the finite model property, therefore completeness with respect to the standard ALPSs follows immediately from the fact that there can be no infinite $>$ chains of more and more plausible worlds.

□

4.3.3 Dynamic axiomatization

We now have to look into the behaviour of rule-applications under ALPSs. Formulas of the form $\langle\rho\rangle\phi$ are interpreted as indicated by the original clause (Definition 4.2.3), only now using the ALPSs corresponding to M and M^ρ.

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

4.3.6. PROPOSITION ($\langle\rho\rangle$-reduction axioms). *The following are valid:*

$$\langle\rho\rangle(z_1s_1 + \ldots + z_ns_n \geq z) \leftrightarrow (cp \geq c_\rho) \wedge A\rho \wedge (z_1s_1^\rho + \ldots + z_ns_n^\rho \geq z)$$

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

Proof:
The claim is easy for the atoms, the Boolean cases, the inequalities, A, $[RAD]_\sigma$, L_K and $L_\Box$. We will only show why the claim holds for I_K and $I_\Box$ because the claims involving K, $\Box$ will then follow from the clause for $\spadesuit$, which is given in terms of $L_\spadesuit$ and $I_\spadesuit$.

- Because of Theorem 4.3.2, it suffices to show that the axiom is valid in an arbitrary RSPM. Let M be such RSPM and w an arbitrary possible world of it.

  ▷ 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:

  $$\text{for all } v \in W^\rho \cap W^I : M^\rho, v \models \phi \qquad (1)$$

  Take arbitrary $u \in W^I \cap W$ and arbitrary $v \in u^\rho$. Then, $v \in W^\rho \cap W^I$, and by (1) and the definitions of V_I 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 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 $u \in W \cap W^I$ such that $v \in u^\rho$. By the truth conditions of $M, w \models I_K[RAD]_\rho\phi$, for all $u \in W^I \cap W$, $M, u \models [RAD]_\rho\phi$, 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_I 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 RSPM and w an arbitrary possible world of it.

  ▷ 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 $W^\rho = \bigcup_{u \in W} u^\rho$:

  $$\text{for all } v \in W^\rho \cap W^I \text{ such that } w \geq^\rho v : M^\rho, v \models \phi \qquad (2)$$

  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 $w \geq^\rho v$, and by (2) and the definitions of V_I 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 $u \in W^I \cap W$ 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 $w \geq^\rho v$, we get $w \geq u$. By the truth conditions of $M, w \models I_\Box[RAD]_\rho\phi$, for all $u \in W^I \cap W$ 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_I and radius, $M^\rho, v \models \phi$ too. Overall, $M^\rho, w \models I_\Box\phi$ and finally $M, w \models \langle\rho\rangle I_\Box\phi$.

□

4.3.7. Theorem (Full axiomatization). *The axiomatic system given by Definition 4.3.3, the reduction axioms of Proposition 4.3.6, with the additional* $[\rho]$*-Necessitation Rule, is sound and complete with respect to ALPSs.*

Proof:
The result follows from Proposition 4.3.6, Theorem 4.3.4, Theorem 4.3.5. □

4.4 Inference and interaction dynamics

In the previous sections, we focused on how deductive reasoning, and the bounds imposed on it by cognitive fatigue, affect the agent's epistemic state. As observed by van Benthem (2008c), 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 by 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. (van Benthem, 2007; Baltag and Smets, 2008b)), fit in our framework, provided that suitable dynamic operators and model transformations are defined.

4.4.1 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 acquisition, 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). Another 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 it.

Extending the syntax. We introduce operators of the form $[\psi!]$ to $\mathcal{L}_{K,\Box}$, such that $[\psi!]\phi$ stands for "after announcing ψ, ϕ is true". We focus on cases where ψ is a propositional formula (i.e. $\psi \in \mathcal{L}_\Phi$, where $\mathcal{L}_\Phi$ is the propositional language based on the set of atoms Φ).[6] Let the language extended with public announcements be called $\mathcal{L}_{PA}$.

[6] 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 Section 4.5.

Extending the semantics. $\mathcal{L}_{PA}$ is interpreted in RSPMs. The new clause concerns public announcements. Semantically, the formula $[\psi!]\phi$ is taken to be true in case: whenever ψ is true, ϕ is true after we eliminate all non-ψ possibilities. This is because the public announcement of ψ is completely trustworthy, so non-ψ worlds, possible or impossible, are not entertained by the agent any more.[7] The other components of the model, namely ord, V_P, V_I, R, are restricted accordingly, while cp does not change, because 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:

4.4.1. Definition (Model transformation by public announcement). Given a RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$, its transformation by the public announcement of $\psi \in \mathcal{L}_\Phi$ is the model $M^{\psi!}$ given by:

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

The conditions on RSPMs 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, Soundness of rules, and Rule availability still hold, because the worlds surviving the announcement still adhere to these restrictions. Succession is preserved because of the way $(W^I)^{\psi!}$ and $V_I^{\psi!}$ are defined; if the conditions of Succession are met in the updated model, the successor world (the expansion) satisfies ψ and is therefore included in $(W^I)^{\psi!}$.

We give the truth clause for public announcements, which follows the standard DEL fashion, only now adapted to the impossible-worlds semantics. In particular:

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

Notice that the formula $[\psi!]\phi$ is vacuously true if ψ 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 ψ are the ones satisfying ψ, possible or not, because arguably *any* non-ψ world will be dropped as a possibility.

With the extended setting, we can bring together external information and the agent's internal reasoning. For instance, suppose that the agent knows $\phi \to \psi$ and has MP available as a rule, and then she comes to know that ϕ from an external

[7] Notice that in the possible-worlds models of DEL, keeping only the ψ-satisfying worlds and deleting all the $\neg\psi$-satisfying worlds have the same effect due to possible worlds being complete. This is not the case in impossible worlds.

source. She may then apply the rule (if affordable) and finally come to know ψ. It is therefore the combination of interaction and internal deductive reasoning that allowed her to know ψ. To illustrate the workings of such combinations, we come back to our bounded version of the Muddy Children Puzzle (Example 4.2.4) and explicitly account for the mix of interaction and inference taking place.

4.4.2. Example (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 depicted in Figure 4.3, Figure 4.4, Figure 4.5, using the same depiction conventions as in Example 4.2.4.

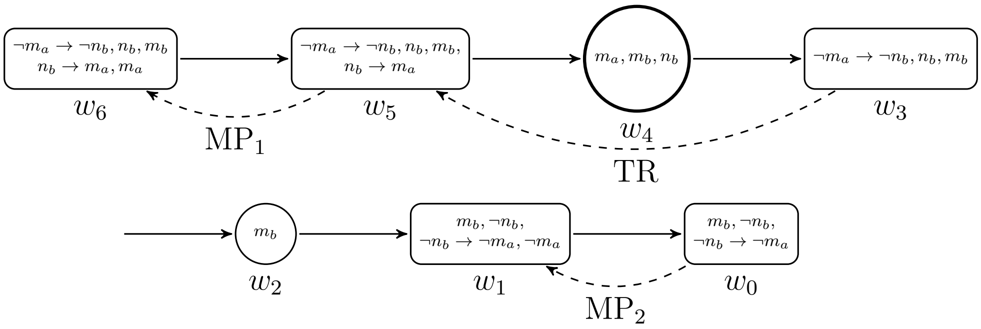


Figure 4.3: The initial model for child a, before the announcement of child b.

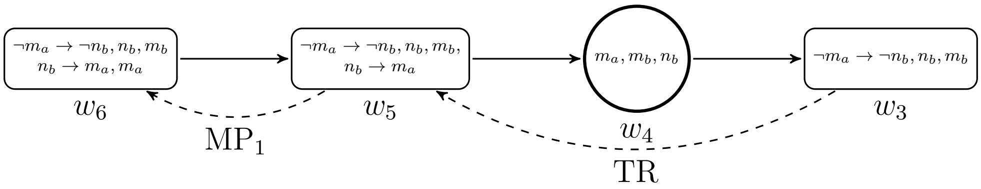

$\neg m_a \to \neg n_b, n_b, m_b$
$n_b \to m_a, m_a$
w_6
$\neg m_a \to \neg n_b, n_b, m_b,$
$n_b \to m_a$
w_5
m_a, m_b, n_b
w_4
$\neg m_a \to \neg n_b, n_b, m_b$
w_3
MP$_1$
TR

Figure 4.4: The model for child a after child b's announcement.

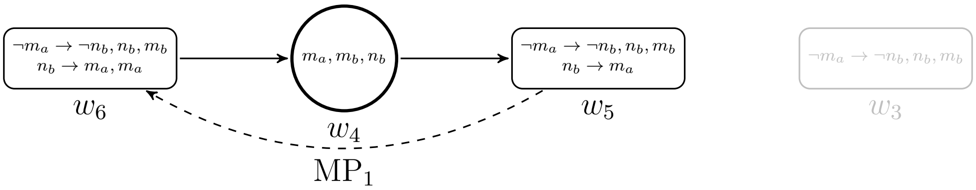

$\neg m_a \to \neg n_b, n_b, m_b$
$n_b \to m_a, m_a$
w_6
m_a, m_b, n_b
w_4
$\neg m_a \to \neg n_b, n_b, m_b$
$n_b \to m_a$
w_5
$\neg m_a \to \neg n_b, n_b, m_b$
w_3
MP$_1$

Figure 4.5: The final model for child a after she performs the TR inference (as in Example 4.2.4), using the information provided by child b.

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. It has been proposed that there are two different kinds of such informational events, presented as “implicit” and “explicit” observations (van Benthem, 2008b,c; Velázquez-Quesada, 2009). In our terms, there can be *effortless* announcements (like the ones defined before) and *effortful* announcements.[8] The latter are just like those in Definition 4.4.1, but they also incur a cost of accepting the announced information. This presupposes 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: $c : \mathcal{L}_R \cup \mathcal{L}_\Phi \to \mathbb{N}^r$. 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; Vigo, 2006; Goodwin and Johnson-Laird, 2013) might assist in determining these costs. The definition of a transformation by an effortful announcement is:

4.4.3. Definition (Model transformation by an effortful public announcement). Take RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp \rangle$. Its transformation by the effortful public announcement of $\psi \in \mathcal{L}_\Phi$ is the model $M^{\psi!}$ given by:

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

The truth clause of an effortful announcement much resembles that of rule-applications. Provided that terms of the form c_ψ are introduced to express the cost of ψ:

$$M, w \models [\psi!]\phi \text{ iff } M, w \models \psi \text{ implies } (M, w \models cp \geq c_\psi \text{ and } M^{\psi!}, w \models \phi)$$

Reduction. In Section 4.3, 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 (without common knowledge) usually involves reduction axioms that allow for replacing formulas with public announcements with – eventually – formulas of the static language (Baltag et al., 1998; van Benthem, 2007; van Ditmarsch et al., 2007; Plaza, 2007). Completeness then follows from the respective complete static base logics. However, the standard reduction axioms would not

[8] This fits well with distinctions 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.

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, precisely 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.[9]

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 4.2.3. Asking that $K\phi$ and $\Box\phi$ are true iff ϕ 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 is not taken for granted.

4.4.2 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*

[9] 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 $[\psi!]I_\spadesuit\phi \leftrightarrow (\psi \rightarrow I_\spadesuit[\psi]\phi)$ 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 $L_\spadesuit$, which behaves normally, that $[\psi!]\spadesuit\phi$ can be reduced. An alternative procedure to this is mentioned in Section 4.5.

(or *lexicographic*) *upgrades* and *conservative upgrades* (Rott, 1989; van Benthem, 2007; Baltag and Smets, 2008b; van Benthem, 2008c, 2011). A radical upgrade with ψ changes the plausibility as follows: ψ-worlds are ranked over the non-ψ worlds but the ranking of worlds within the two zones remains intact. Regarding conservative upgrades, the most plausible of the ψ-worlds are ranked over all other worlds and the rest remain unchanged.[10] In 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 $\psi \in \mathcal{L}_\Phi$, to denote radical upgrades with ψ. More specifically, $[\psi \Uparrow]\phi$ reads "after the radical upgrade with ψ, ϕ is true".

Extending the semantics. We present the model transformation by a radical upgrade – again, assuming it amounts to an effortless process.[11] As an auxiliary step take: $\geq^{\psi\Uparrow} = (\geq \cap (W \times [[\psi]])) \cup (\geq \cap (\overline{[[\psi]]} \times W)) \cup (\sim \cap (\overline{[[\psi]]} \times [[\psi]]))$, where $\overline{[[\psi]]}$ denotes the set of worlds where ψ is not satisfied. Then:

4.4.4. Definition (Model transformation by a radical upgrade). Take RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp \rangle$. Its transformation by a radical upgrade with $\psi \in \mathcal{L}_\Phi$ is the model $M^{\psi\Uparrow} = \langle W^P, W^I, W, ord^{\psi\Uparrow}, V_P, V_I, R, cp \rangle$, where $ord^{\psi\Uparrow}$ is any function from the set $\{f : W \to \Omega \mid \text{ for any } w, u \in W : f(w) \geq f(u) \text{ iff } w \geq^{\psi\Uparrow} u\}$.

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 ψ-worlds over non-ψ ones. The properties of our models are clearly preserved. Then 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 given by 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).[12]

4.4.5. Example (Bounded muddy children and radical upgrades). We now approach the scenario of Example 4.4.2 using the "softer" version of the Puzzle. That is, we take child b as a reliable, but not infallible, source of information.

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

[11] Defining *effortful* upgrades, in the spirit explained before, is also possible.

[12] 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.

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 (Figure 4.6).

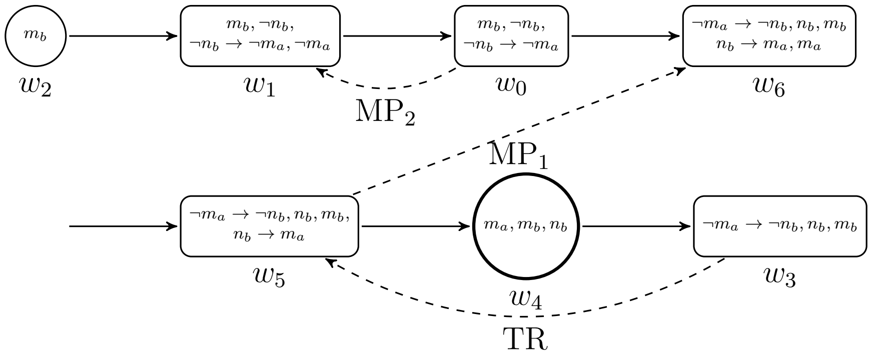


Figure 4.6: The model of Figure 4.3 after the upgrade with n_b. Clearly, $\Box n_b$ is satisfied at the actual world (w_4), unlike Kn_b. Provided that TR and MP_1 are affordable and available, $\langle TR\rangle\langle MP_1\rangle\Box m_a$ is also satisfied, unlike $\langle TR\rangle\langle MP_1\rangle Km_a$.

4.4.3 Rule dynamics: learning and forgetting rules

The rules that are available to the agent have been assumed to be fixed. However, it can well be that agents *learn* and *forget* rules, hence performing better or worse in reasoning tasks. To that end, one can introduce dynamic operators for learning and forgetting rules, and corresponding model transformations that modify the R function accordingly. For example, it has been argued that MT, unlike MP, is not a primitive rule (Rips, 1994), so it might be that the rule is not even available to the agent, who needs to *learn* it and then apply it.

Extending the syntax. We introduce dynamic operators of the form $\langle +\rho\rangle$ (resp. $\langle -\rho\rangle$), such that: $\langle +\rho\rangle\phi$ (resp. $\langle -\rho\rangle$) says "after the agent learns (resp. forgets) ρ, ϕ is true".

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

4.4.6. Definition (Model transformation by learning a rule). Given a RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$, its transformation by learning a rule ρ is a model $M^{+\rho} := \langle (W^P)^{+\rho}, W^I, W^{+\rho}, ord^{+\rho}, V_P^{+\rho}, V_I, R^{+\rho}, cp\rangle$ with (i) $(W^P)^{+\rho} = \{w \in W^P \mid M, w \models tr(\rho)\}$, (ii) $W^{+\rho} = (W^P)^{+\rho} \cup (W \cap W^I)$, (iii) $R^{+\rho}(w) = R(w) \cup \{\rho\}$, while (iv) $ord^{+\rho} = ord_{|_{W^{+\rho}}}$ and $V_P^{+\rho} = V_{P|_{(W^P)^{+\rho}}}$.

4.4.7. Definition (Model transformation by forgetting a rule). Given RSPM $M = \langle W^P, W^I, W, ord, V_P, V_I, R, cp\rangle$, its transformation by forgetting a rule ρ is $M^{-\rho} := \langle W^P, W^I, W, ord, V_P, V_I, R^{-\rho}, cp\rangle$ with $R^{-\rho}(w) = R(w) \setminus \{\rho\}$.

The model transformations clearly preserve the properties of RSPMs. The clauses for learning/forgetting rules at $w \in W^P$ are given by:

$$M, w \models \langle +\rho \rangle \phi \text{ iff } M, w \models tr(\phi) \text{ and } M^{+\rho}, w \models \phi$$
$$M, w \models \langle -\rho \rangle \phi \text{ iff } M^{-\rho}, w \models \phi$$

Reduction. We can then develop the method of Section 4.3 and extract a sound and complete axiomatization for the extended language via the reduction axioms for actions of learning/forgetting rules:

$$\begin{array}{l|l}
\langle +\rho \rangle p \leftrightarrow tr(\rho) \wedge p & \langle +\rho \rangle \neg \phi \leftrightarrow tr(\rho) \wedge \neg \langle +\rho \rangle \phi \\
\langle +\rho \rangle (\phi \wedge \psi) \leftrightarrow \langle +\rho \rangle \phi \wedge \langle +\rho \rangle \psi & \langle +\rho \rangle L_{\spadesuit} \phi \leftrightarrow tr(\rho) \wedge L_{\spadesuit} \langle +\rho \rangle \phi \\
\langle +\rho \rangle I_{\spadesuit} \phi \leftrightarrow tr(\rho) \wedge I_{\spadesuit} \phi & \langle +\rho \rangle \spadesuit \phi \leftrightarrow \langle +\rho \rangle L_{\spadesuit} \phi \wedge \langle +\rho \rangle I_{\spadesuit} \phi \\
\langle +\rho \rangle A\sigma \leftrightarrow tr(\rho) \wedge A\sigma, \text{ for } \sigma \neq \rho & \langle +\rho \rangle A\rho \leftrightarrow tr(\rho) \wedge \top \\
\langle +\rho \rangle (z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow tr(\rho) \wedge (z_1 s_1 + \ldots + z_n s_n \geq z) & \langle +\rho \rangle [RAD]_{\sigma} \phi \leftrightarrow [RAD]_{\sigma} \langle +\rho \rangle \phi
\end{array}$$

$$\begin{array}{l|l}
\langle -\rho \rangle p \leftrightarrow p & \langle -\rho \rangle \neg \phi \leftrightarrow \neg \langle -\rho \rangle \phi \\
\langle -\rho \rangle (\phi \wedge \psi) \leftrightarrow \langle -\rho \rangle \phi \wedge \langle -\rho \rangle \psi & \langle -\rho \rangle L_{\spadesuit} \phi \leftrightarrow L_{\spadesuit} \langle -\rho \rangle \phi \\
\langle -\rho \rangle I_{\spadesuit} \phi \leftrightarrow I_{\spadesuit} \phi & \langle -\rho \rangle \spadesuit \phi \leftrightarrow \langle -\rho \rangle L_{\spadesuit} \phi \wedge \langle -\rho \rangle I_{\spadesuit} \phi \\
\langle -\rho \rangle A\sigma \leftrightarrow A\sigma, \text{ for } \sigma \neq \rho & \langle -\rho \rangle A\rho \leftrightarrow \bot \\
\langle -\rho \rangle (z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow (z_1 s_1 + \ldots + z_n s_n \geq z) & \langle -\rho \rangle [RAD]_{\sigma} \phi \leftrightarrow [RAD]_{\sigma} \langle -\rho \rangle \phi
\end{array}$$

4.5 Conclusions

By combining DEL and plausibility impossible-wolds semantics, we modelled an agent who is fallible but boundedly rational in a framework capable of representing a variety of attitudes and actions. It was shown that our models can be reduced to syntactic, possible-worlds, plausibility structures that allow 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.

We therefore showed that the resource-sensitive semantics, initially presented in Chapter 3, is still compatible with the progress that has been made in the DEL literature, while free from idealizations we argued against. This flexibility is also demonstrated in its technical features. In the next chapter, we will investigate the implications of such logical settings in the dual process theories of reasoning.

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 envisage a principle of moderation, achieved via the introduction of effortful introspective rules whose semantic effect is similarly projected on the structure of the model. This will be pursued in Part III.

Moreover, it is interesting to search for alternatives to the use of special operators in providing reduction axioms for rule-applications and announcements. This is 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). Action models with postconditions (van Benthem et al., 2006), along with the *set-expressions* used by Velázquez-Quesada (2014) to embed these into awareness frameworks, can help in obtaining reduction axioms for both rule applications and communicative actions, through a unified treatment. This will also be pursued in Part III.

Chapter 5

The logic of fast and slow thinking

The diversity of mental processes has been a key component of the alternative rationality picture, which we seek to formalize, as explained in Chapter 2. In this chapter, we further motivate the logical modelling of aspects of the dual process theories of reasoning and we show that our constructions for resource-bounded reasoning can be utilized in this context.[1]

In particular, we present a framework for epistemic logic, modelling the logical aspects of System 1 ("fast") and System 2 ("slow") cognitive processes, as per dual process theories of reasoning. The framework combines impossible-worlds semantics with the techniques of DEL. It models non-logically-omniscient, but moderately rational agents: their System 1 makes fast sense of incoming information by integrating it on the basis of their background knowledge and beliefs. Their System 2 allows them to slowly, stepwise unpack some of the logical consequences of such knowledge and beliefs, by paying a cognitive cost. The framework is applied to three instances of limited rationality, widely discussed in cognitive psychology: Stereotypical Thinking, the Framing Effect, and the Anchoring Effect.

We further motivate this attempt in Section 5.1. We then briefly recap the distinction between the two systems and explain the sense in which we claim to *logically* model it in Section 5.2. These observations will serve as the background for our logical model of the two Systems, presented in Section 5.3. In Section 5.4, the framework is put to work in the modelling of the three case studies. We close this chapter, and Part II in general, with a philosophical coda in Section 5.5, where we wonder whether the logical constructions so far discussed are normative. We answer that they are, but their rational "ought", unlike the "ought" of normal, static epistemic logics, implies "can".

[1] The chapter is based on Solaki et al. (2019).

5.1 Econs, Logons, and Humans

The gap between human and idealized agents, and the argumentation towards an alternative rationality picture do not only challenge the ventures of logicians, but of formal modellers in general. For example, 2017 Nobel laureate in economics Richard Thaler dubbed "Econs" and "Humans" two different species studied, respectively, by mainstream economists and by behavioral and cognitive scientists (Thaler and Sunstein, 2008). Econs are the agents of classical economic theory: fully consistent, endowed with stable and well-ordered preferences as per Bernoulli's expected utility theory. Of course, the terminology implies that Humans, unlike Econs, are the real thing. The discrepancies between the two kinds are no different to the ones that have sparked the Rationality Debate, discussed in Chapter 2. As 2002 Nobel laureate in economics Daniel Kahneman has it:

> [Assume] rationality is logical coherence – reasonable or not. Econs are rational by this definition, but there is overwhelming evidence that Humans cannot be. [...] The definition of rationality as coherence is impossibly restrictive; it demands adherence to rules of logic that a finite mind is not able to implement. Reasonable people cannot be rational by that definition, but they should not be branded as irrational for that reason. (Kahneman, 2011, p. 411)

Now just as mainstream economics has forgotten Humans in exchange for a focus on Econs, so has mainstream logic forgotten them to focus on *Logons*. Adopting this parlance to describe the contrast, we name this way the logically omniscient agents studied in EL. In fact, Econs may just be Logons engaged in rational choice. The focus on Logons has opened a rift between logic and cognition, similar to the one between the latter and economics. We explained why we think that such a conclusion has been distorted by a misconception on the role of "logic". The goal of this chapter is to present a model that does more justice to Humans by modelling the logical aspects of the distinction between System 1 and System 2 (or, in Kahneman's more colourful terminology, fast and slow thinking), which has played a key role in the Rationality Debate and motivated our alternative picture.

5.2 Background

The talk of System 1 and System 2 had a key role in countering the picture of agents as Econs in economics. We argued that it has a key role in countering the picture of agents as Logons in EL. Systematic errors in reasoning and choice are not to be taken as corruption of rationality. Rather, they are grounded in the ordinary workings of the machinery of cognition – specifically, in a combination of mistakes due to System 1 (which, however, often conforms to logic (Bago and

Neys, 2017)) and System 2 (which can run out of cognitive resources, or be lazy when it should take over from System 1).

Recall that the operations of System 1 are supposed to be fast, automatic, and associative, governed by habit, biases, and heuristics. They typically have no cognitive cost. System 1's task is to make sense of the continuously incoming new information, integrating it with our background beliefs and building, on their basis, a coherent picture starting from minimal clues (Paul is French: does he like red wine?). In Kahneman's words:

> The main function of System 1 is to maintain and update a model of your personal world, which represents what is normal in it. [...] System 1 excels at constructing the best possible story that incorporates ideas currently activated, but it does not (cannot) allow for information it does not have. (Kahneman, 2011, p. 71 and p. 85)

The operations of System 2 are slower, stepwise, rule-based, deliberately controlled, and have cognitive costs (What is 19×26?). System 2 exploits the workings of System 1 to generate its own outputs, following an orderly application of steps:

> I describe System 1 as effortlessly originating impressions and feelings that are the main sources of the explicit beliefs and deliberate choices of System 2. The automatic operations of System 1 generate surprisingly complex patterns of ideas, but only the slower System 2 can construct thoughts in an orderly series of steps. (Kahneman, 2011, p. 21)

When System 2 takes over, it engages in reasoning processes, of which deductive reasoning is a key example, based on available information. Its slow, stepwise and rule-adhering workings generate our – now explicit – knowledge and beliefs. To unpack information, System 2 breaks larger tasks into parts:

> We normally avoid mental overload by dividing our tasks into multiple easy steps, committing intermediate results to long-term memory or to paper rather than to an easily overloaded working memory. We cover long distances by taking our time and conduct our mental lives by the law of least effort. (Kahneman, 2011, p. 38)

Given that the process is effortful, and our resources are bounded, it must eventually halt, whether it succeeds or not. This is in accordance with our experience of occasionally failing in demanding tasks due to cognitive overload.

As clarified by Evans (2018), one should not take System 1 as merely descriptively representing what people, as a matter of fact, do most of the time, and System 2 as embedding the normative standards of rationality. On the contrary, System 2 can occasionally fail to do its job in correcting the mistaken outputs of

System 1, which, on the other hand, can display good logical intuitions and get things right on most occasions (Bago and Neys, 2017).

Dual process theories have been mostly neglected by formal modelers in logic (relevant exceptions are Stenning and van Lambalgen (2008); Balbiani et al. (2019)). We aim to contribute to filling the gap by modelling the logical aspects of System 1 and System 2 reasoning activities: those that are connected to logical inferences – a most classical topic of logical investigation – and the formation and revision of beliefs – a core topic of doxastic-epistemic logics and belief revision theory.[2] The ways the two cognitive systems shape our epistemic/doxastic state will be expressed as model-changing actions on appropriate plausibility models, in the lines of Chapter 4.

5.3 Modelling fast and slow thinking

In this section, we introduce a new logical system to represent and model agents capable of fast and slow thinking. We break down our aims as follows:

- ▷ Enrich the standard possible-worlds semantics of EL with impossible worlds to encode the beliefs of a human, logically competent but not omniscient, agent.
- ▷ Use tools from DEL to model how incoming information is automatically incorporated by System 1 into the currently held beliefs.
- ▷ Use tools from DEL to capture the agent's stepwise deductive reasoning via System 2.
- ▷ Allow for the interaction of the two systems.
- ▷ Account for how the two systems differ in terms of cognitive resource consumption.

5.3.1 Syntax

We develop an epistemic-doxastic language that also represents the workings of the two systems. In particular, we use operators for *belief* (B), and *(defeasible) knowledge* ($\Box$). Our focus on this more graded outlook of attitudes, incorporating both belief and a weaker notion of knowledge, is better aligned with the cognitive workings of System 1 and System 2, and experimental results thereof (e.g. on belief bias). The language also has dynamic operators to express (1) System 1's fast upgrades in the arrangement of our beliefs – policies of automatic integration

[2] According to the official dual process theories, the two systems engage in a range of further activities: System 1, for instance, deals with face recognition, orientation, perception, etc. System 2 deals with probabilistic estimates, the weighing of options, etc. An expansion of the model proposed below in the direction of probabilistic reasoning may be especially interesting, as dual process theories have been developed in relation to the new Bayesian approaches in the psychology of reasoning (Elqayam, 2018); this is discussed in Chapter 9.

of new information, and (2) System 2's cognitively costly choices and applications of logical inference rule-schemes.

5.3.1. Definition (Dual process language). Given a set Φ of propositional atoms and a set of inference rules R available to the agent, the *dual process language* $\mathcal{L}_{\mathsf{DP}}$ is given by:

$$\phi \quad ::= \quad p \quad | \quad \neg\phi \quad | \quad \phi \wedge \phi \quad | \quad \Box\phi \quad | \quad B\phi \quad | \quad \alpha\, \phi$$

such that:

- $p \in \Phi$
- $\Box\phi$ reads "the agent defeasibly knows that ϕ".
- $B\phi$ reads "the agent believes that ϕ".
- α is schematic for a model-changing action performed in thought. These can be of the two aforementioned kinds:
  (1) $[\psi \Uparrow]$, where ψ is a propositional formula, denotes a fast upgrade with ψ: given incoming information ψ, the agent automatically makes plausible sense of the situation in light of her background knowledge and beliefs. Then $[\psi \Uparrow]\phi$ reads "after upgrading with ψ, ϕ is true".
  (2) $\langle\rho\rangle$, where $\rho \in R$, that is, an inference rule available to the agent. The agent can deliberately choose one of them, apply it to some available information and, as we shall see, pay some cognitive cost for it. Then $\langle\rho\rangle\phi$ reads "after some application of inference rule ρ, ϕ is true".

One can, in principle, build dynamic models with rules representing various kinds of rule-based System 2 reasoning. For the purposes of this chapter, however, we will take R as comprising just schemes of rules of elementary logic, such as *Modus Ponens* or *Conjunction Introduction*.[3] We should note that, as clarified by recent literature (Bago and Neys, 2017; Ball and Thompson, 2018), System 1 is also capable of detecting and appreciating simple logical forms. The key twofold difference between System 1 and System 2 in this respect is that the latter, but not the former, can *choose* which logical rules to apply, and must *pay* a cognitive cost for it. This is why we focus on schemes of rules and emphasize that the applications of System 2 are deliberate, thus highlighting the differences between the two systems.[4]

[3] The idea of such operators comes from Rasmussen (2015); Bjerring and Skipper (2018), themselves drawing on Duc (1997).

[4] The inequalities appearing in Definition 3.2.3, Definition 4.2.1 could also be introduced in this language to obtain the full technical treatment of the previous chapters (Chapter 3, Chapter 4). We now focus on illustrating the modelling of fast and slow processes within the context of the constructions of Part II. The dependence on resources will be embedded into the workings of the model transformations and will not have to be reflected in the syntax.

5.3.2 Semantics

In what follows, we employ a plausibility model, similar to the one of Chapter 4, to capture the workings of the two systems. We impose a plausibility ordering on worlds, encoding the agent's background beliefs: the more plausible a world looks given the agent's experience, biases, etc., the better it is ranked (the ordering is qualitative, mirroring belief entrenchment). Plausibility is instrumental in modelling, as we will see, the changes induced by both (1) the fast incorporation of external information by System 1, (2) the slow reasoning processes of System 2.

We need ways to represent which cognitive resources are explicitly depleted during System 2 reasoning (time, memory, etc.), what each reasoning step costs, and what the agent can afford with respect to them. Each step corresponds to an application of an inference rule. Yet not all inference rules require equal cognitive effort, as indicated by experimental evidence. This is why we again fix Res, a finite set of resources, such as *memory*, *time*, etc, and take $r := |Res|$, i.e. the number of resources. We also fix R, the set of inference rules available to the agent; simplifying the constructions of Chapter 4, we will not incorporate modifications on the set of rules. Finally, we introduce the *cognitive cost function* $c : R \to \mathbb{N}^r$, which is such that every inference rule $\rho \in R$ is assigned a particular cost with respect to each resource.[5]

5.3.2. Definition (Dual process plausibility model). Given Res, R and c, a *dual process plausibility model* (DPPM) is a tuple $M = \langle W^P, W^I, ord, V_P, V_I, cp\rangle$ where:

- W^P, W^I are non-empty sets of possible and impossible worlds respectively.
- $ord : W \to \Omega$ is a function from $W := (W^P \cup W^I)$ to the class of ordinals Ω, assigning one to each world. Intuitively: the smaller the ordinal is, the more plausible the world.
- $V_P : W^P \to \mathcal{P}(\Phi)$ is a function that assigns to each $w \in W^P$ the set of atomic formulas true at w.
- $V_I : W^I \to \mathcal{P}(\mathcal{L}_{\mathsf{DP}})$ is a function assigning to each impossible world in W^I a set of formulas in $\mathcal{L}_{\mathsf{DP}}$. It assigns to each $w \in W^I$ *all* formulas, atomic or complex, true at w.[6] Thus, V_I maps logically complex formulas to truth values directly at impossible worlds, in a non-recursive fashion: this allows such worlds to break any (non-trivial, i.e. different from 'If ϕ, then ϕ') logical principle. However, in accord with Minimal Consistency, which we adopt on the same grounds as in the previous chapters, we stipulate that $\{\phi, \neg\phi\} \nsubseteq V_I(w)$ for all $w \in W^I$.

[5] The cost function captures the differing cognitive difficulty of rule schemes, relative to one another (recall the ordering of inferences envisaged by Cherniak (1986)), without additionally accounting for the differing complexity of their inputs.

[6] The worlds are considered to be valuation-wise unique, i.e. the valuation functions taking care of possible and impossible worlds are injective. This is to serve simplicity: we avoid a multiplicity of worlds unnecessary for our modelling purposes.

- cp denotes the agent's cognitive capacity, i.e. $cp \in \mathbb{N}^r$, intuitively standing for what the agent is able to afford with respect to each resource.

A *pointed* DPPM consists of a DPPM M and a designated world of it. The function ord extracts a plausibility ordering in the usual sense, i.e. a binary relation on W: $w \geq u$ iff $ord(w) \geq ord(u)$. The ranking of worlds is reflected in the ordering of ordinals. The intended reading is "w is no more plausible than u". The ordering satisfies reflexivity, transitivity, connectedness, and converse wellfoundedness. Fast and slow thinking will be reflected in the interpretation of the formulas involving the operators for upgrades and inference rule applications. We thus have to define how the model changes through these actions.

5.3.3 Model transformations, fast and slow

The fast updater

Each transformation is governed by its corresponding system: thus, System 1's actions of integrating new information will be affected by the agent's stereotypes, biases, experience, etc., as these are hardwired in the initial plausibility ordering. Based on this, the system incorporates new information by prioritizing the worlds satisfying it. That is, an upgrade with ψ changes the plausibility ordering as follows: ψ-worlds become more plausible than non-ψ ones (i.e. those that do not satisfy ψ) keeping the previous ordering intact within the two zones. Moreover, as fast thinking, this activity requires no effort; therefore the relevant components of the model should be unaffected by the upgrade.

5.3.3. Definition (Transformation by a System 1 upgrade). Given a DPPM $M = \langle W^P, W^I, ord, V_P, V_I, cp \rangle$, its transformation by $\psi \Uparrow$ is a model $M^{\psi\Uparrow} = \langle W^P, W^I, ord^{\psi\Uparrow}, V_P, V_I, cp \rangle$ where $ord^{\psi\Uparrow}$ can be every function from the set $\{f : W \to \Omega \mid \text{ for any } w, u \in W : f(w) \geq f(u) \text{ iff } w \geq^{\psi\Uparrow} u\}$.[7]

The characterization via ordinals does not interfere with radical upgrades. We will not be interested in the assigned number *per se*, but in the action-induced re-arrangement (i.e. plausibility of worlds relative to other worlds). Thus, all functions from $\{f : W \to \Omega \mid \text{ for any } w, u \in W : f(w) \geq f(u) \text{ iff } w \geq^{\psi\Uparrow} u\}$ work for our purposes. The properties of DPPMs are clearly preserved.

[7] To determine $ord^{\psi\Uparrow}$, first consider the relation $\geq$ that can be derived from it. As an auxiliary step take: $\geq^{\psi\Uparrow} = (\geq \cap (W \times [[\psi]])) \cup (\geq \cap (\overline{[[\psi]]} \times W)) \cup (\sim \cap (\overline{[[\psi]]} \times [[\psi]]))$, that is the familiar re-arrangement due to a radical upgrade as found in DEL and Definition 4.4.4.

The slow controller

We account for the stepwise, deliberate, and cognitively costly workings of System 2 via our rule-application operators. To define the transformation induced by these operators, we employ the notion of *ρ-accessibility.* For a pointed model (M', w) to be *ρ-accessible* from a given pointed DPPM (M, w), the set $P_{\geq}(w) := \{u \in W \mid w \geq u\}$ of worlds at least as plausible as w is replaced by a choice of worlds reachable by an application of ρ from the elements of $P_{\geq}(w)$, while the remaining ordering is adapted accordingly. We focus on the more or equally plausible worlds, as these would be prioritized whenever one applies an inference rule.

By specifying the effect of each rule separately, it is possible to trace back a sequence of slow reasoning, unravel it and verify its order-sensitivity. In addition, the agent's cognitive capacity should be reduced by the cost of applying this particular inference step.

To capture the change induced by applications of inference rules, we first have to encode their effect on the structure of our models. The effect of applying a rule is an expansion of the agent's factual information. We first introduce the following, assuming that propositional formulas are assessed as usual in possible worlds:

5.3.4. Definition (Propositional truths). Let M be a DPPM, $w \in W$ a world of the model, and $\mathcal{L}_\Phi$ the standard propositional language based on Φ. The set of *propositional truths* for w is given by: $V^*(w) = \begin{cases} \{\phi \in \mathcal{L}_\Phi \mid M, w \models \phi\} \text{ if } w \in W^P \\ \{\phi \in \mathcal{L}_\Phi \mid \phi \in V_I(w)\} \text{ if } w \in W^I \end{cases}$

That is, V^* is in fact determined by V_P and V_I. Next, we take ρ_k, a particular instance of the inference rule ρ. This has a set of (propositional) premises, denoted by $pr(\rho_k)$, and a conclusion, denoted by $con(\rho_k)$. We can then impose the condition of Succession again:

For every $w \in W$, if (1) $pr(\rho_k) \subseteq V^*(w)$, (2) $\neg con(\rho_k) \notin V^*(w)$ and (3) $con(\rho_k) \neq \neg\phi$ for all $\phi \in V^*(w)$, then there is $u \in W$ such that $V^*(u) = V^*(w) \cup \{con(\rho_k)\}$.

We use $V^*(w) \vdash_\rho V^*(u)$ to say that for some instance of ρ, u *expands* w in terms of this condition. If $pr(\rho_k) \subseteq V^*(w)$ for no instance ρ_k of ρ, we take the only ρ-expansion of w to be itself. If $pr(\rho_k) \subseteq V^*(w)$ for an instance ρ_k of ρ, but condition 2 or 3 is violated, then there is simply no ρ-expansion with regard to this instance; this is instrumental in preserving Minimal Consistency. Notice that by conjoining successive rules, such as $\rho_1, \ldots, \rho_n$, the notation can be generalized to $\vdash_{\rho_1,\ldots,\rho_n}$.

5.3.5. Definition (Rule-specific radius). Given an inference rule $\rho \in R$, the ρ-radius of a world $w \in W$ is $w^\rho = \{u \mid V^*(w) \vdash_\rho V^*(u)\}$.

A member of w^ρ is therefore a ρ-expansion of w. Under the conditions, $\vdash_\rho$ is such that $V^*(u)$ preserves $V^*(w)$ and extends it just by a conclusion of ρ_k. In addition, we have a monotonicity feature: ρ-expansions (as per the name) enrich the state from which they originate, in terms of ρ: inferences are not defeated as reasoning steps are taken, but only to the extent that Minimal Consistency is respected. Note that $w^\rho = \{w\}$ for $w \in W^P$ due to the deductive closure of possible worlds, while the ρ-radius of impossible worlds can contain different ρ-expansions.

Not all instances of a rule are equally informative. Compare an application of *Conjunction Introduction* that allows the agent to conclude that $\phi \wedge \psi$, from ϕ and ψ, and an application that generates $\phi \wedge \phi$ from ϕ. Rips (1994) classifies rules into *self-constraining* and *self-promoting*. Self-constraining rules, such as *Modus Ponens*, generate a limited number of new sentences from their premises. Self-promoting rules, such as *Conjunction Introduction*, generate an infinite number of conclusions from their premises. It is natural to aim at reducing the space W^I from the (possibly infinite) worlds corresponding to non-informative applications of self-promoting rules. This is not to say that the conclusions of these applications should not be available to the agent. In principle, the setting should allow for applications leading to the agent knowing/believing such conclusions. In order to do justice to both points, the modeller might simply assume that a world's expansion corresponding to a non-informative instance is the world itself. However, we abstain from imposing this as a strict condition on the general class of our models, in order to allow for the modelling of a variety of types of agents that may require different readings of informativeness, thus different compositions of a world's radius. To capture the *choice* the agent's System 2 exercises, we define:

5.3.6. Definition (Choice function). Let $\mathcal{C} : \mathcal{P}(\mathcal{P}(W)) \to \mathcal{P}(\mathcal{P}(W))$ be a choice function that takes a set $\mathcal{W} = \{W_1, \dots, W_n\}$ of sets of worlds as input and returns the set $\mathcal{C}(\mathcal{W})$ of sets of worlds which results from all the ways in which exactly one element can be picked from each non-empty $W_i \in \mathcal{W}$. A member of $\mathcal{C}(\mathcal{W})$ is called a choice of $\mathcal{W}$.

A choice function on a set consisting of the radii of worlds will capture how System 2 can deliberate and choose its next step of slow thinking. Given the aforementioned remark on informative and non-informative instances, the several choices that the function yields correspond to the different effects of applying a particular rule.

Now we can explain the effect of System 2's applications of an inference rule ρ: if a world u was considered at least as plausible as the point w of a pointed DPPM before an application of the rule ρ, but does not survive such application, then the agent can rule u out as a doxastic or epistemic possibility. This world must have been an impossible world: a possible world will always survive applications of inference rules, as its radius amounts to itself. What was taken as an epistemic possibility has been spotted as impossible by a slow computation of System 2. Once we rule out such worlds, we preserve the previous ordering to the extent that it is unaffected by the application of the inference rule. That is, there might be parts of the model still independent of this particular application of deductive reasoning, remaining influenced by System 1 alone.

To make this precise, we use the ordinal function and the notion of rule-specific radius. Let $M = \langle W^P, W^I, ord, V_P, V_I, cp\rangle$ be a DPPM. We spell out the transformation in steps:

Step 1 Let (M, w) be M pointed at w. Then, given an inference rule ρ, take $P^\rho(w) := \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^\rho \mid v \in P_{\geq}(w)\})$. In words, a choice of ρ-expansions of the worlds initially considered at least as plausible as w.

Step 2 Based on the argument used above, if $u \in P_{\geq}(w)$ but $u \notin P^\rho(w)$, then u must be excluded from the new model. So in any case, the ρ-accessible pointed model (M', w) should be such that its set of worlds is $W^\rho = W \backslash \{u \in P_{\geq}(w) \mid u \notin P^\rho(w)\}$. The elimination in fact affects W^I.

Step 3 We now develop the new ordering ord^ρ following the application of the inference rule. Let $u \in W^\rho$:

1. If $u \notin P_{\geq}(w) \cup P^\rho(w)$, then $ord^\rho(u) = ord(u)$, i.e. the assigned ranking remains the same, for worlds that were less plausible than w and are not contained in the choice.
2. Next consider $u \in P^\rho(w)$. This means that there is at least one $v \in P_{\geq}(w)$ such that $u \in v^\rho$ for the particular choice C that gave rise to $P^\rho(w)$. Denote the set of such v's by T. Then $ord^\rho(u) = ord(z)$ for $z \in min(T)$. Therefore, if a world is in $P^\rho(w)$, then it takes the position of the most plausible of the worlds from which it originated.

Step 4 Finally, for worlds $u, v \in W^\rho$: $u \geq^\rho v$ iff $ord^\rho(u) \geq ord^\rho(v)$, therefore again all the required properties of the ordering are preserved.

Step 5 The other components remain unchanged, except from V_P, V_I which are restricted to the surviving worlds, and $cp^\rho := cp - c(\rho)$. Reducing the value of cognitive capacity models slow thinking as resource-consuming.[8] This also determines whether the transformed model is ρ-accessible, i.e. whether the resulting capacity is entirely depleted or not (recall that $cp \in \mathbb{N}^r$).

[8] Agents can, of course, use methods like note-taking, or resort to other external devices, for the offloading of cognitive resources such as memory. In terms of our quantitative assignments, this would entail an increase in capacity. This can be easily achieved by the introduction of actions that increase the value of cp. It does not affect the crucial aspect hereby captured: the resource-consumption caused by System 2.

Here is an example to get a feel of how this model transformation works:

5.3.7. EXAMPLE. Let s stand for "the survival rate within one month of the surgery is 90%", m for "the mortality rate within one month of the surgery is 10%", r for "the surgery is safe". Suppose Jill entertains the worlds depicted in the DPPM M of Figure 5.1, where $W^P = \{w_1\}$ and $W^I = \{w_2, w_0\}$. Take $ord(w_2) = 2, ord(w_1) = 1, ord(w_0) = 0$.

We follow the conventions of Chapter 4 in depicting DPPMs. For the possible world w_1, we list only the propositional atoms it satisfies, since all the rest can be computed recursively. For the impossible worlds, we write down all the propositional formulas satisfied there (and only those) to illustrate Succession and the definitions involved in the model transformation.

All worlds validate $s \to r$, s, r and $s \to m$, but m does not hold in the most plausible world w_0: the most plausible world is such to represent that Jill has not inferred that m follows from s[9] although she has inferred r from s. Finally, focusing on the resources of time and memory (as in Example 3.2.9 and Example 4.2.4), we take the cost of applying *Modus Ponens* to be $c(MP) = (1, 2)$, and the capacity of the agent to be $cp = (15, 7)$.

We then unravel step-by-step the model transformations due to applications of MP (once we give our semantic clauses, we will see how these transformations affect the development of Jill's epistemic and doxastic state). In search of all the ways the pointed model (M, w_1) can change following an application of the rule MP, we follow the procedure sketched above:

Step 1 First, we compute $\{v^{MP} \mid v \in P_{\geq}(w_1)\}$. It amounts to $\{\{w_1\}, \{w_0, w_2\}\}$.

- As a result, $\mathcal{C}(\{\{w_1\}, \{w_0, w_2\}\}) = \{\{w_1, w_0\}, \{w_1, w_2\}\}$.

  So $P^{MP}(w_1) = \{w_1, w_0\}$ or $P^{MP}(w_1) = \{w_1, w_2\}$.

- 1. In case $P^{MP}(w_1) = \{w_1, w_0\}$:

     Step 2 $W^{MP} = W$

     Step 3 Since $w_2 \notin P^{MP}(w_1) \cup P_{\geq}(w_1)$, $ord^{MP}(w_2) = ord(w_2) = 2$. Next, $w_1 \in P^{MP}(w_1)$ and $w_1 \in w_1^{MP}$, so $ord^{MP}(w_1) = ord(w_1) = 1$. Finally $w_0 \in P^{MP}(w_1)$ and $w_0 \in w_0^{MP}$, so $ord^{MP}(w_0) = ord(w_0) = 0$.

     The MP-transformed model is in this case identified with the initial model because it was generated by an application of MP that yielded no new information.

  2. In case $P^{MP}(w_1) = \{w_1, w_2\}$:

     Step 2 $W^{MP} = W \setminus \{u \in \{w_1, w_0\} \mid u \notin \{w_1, w_2\}\} = \{w_1, w_2\}$.

[9] This is just an example of *framing* as discussed by Kahneman (2011). More specifically, it has been shown that subjects are risk-averse when an option is presented in terms of gains and risk-seeking when presented in terms of losses.

Step 3 As above, $ord^{MP}(w_1) = ord(w_1) = 1$. Then, $w_2 \in P^{MP}(w_1)$ and, checking from which world(s) it originated in the particular choice, we find $w_2 \in w_0^{MP}$, so $ord^{MP}(w_2) = ord(w_0) = 0$.

The MP-transformed model is in this case different; the impossible world that did not satisfy m, despite satisfying both $s \to m$ and s, was uncovered by Jill, precisely because she used an application of MP that generated new information. The effect of taking this slow inferential step is now reflected in the new model.

Step 4 The new plausibility ordering is depicted in the figure.

Step 5 The new valuation is obviously restricted to the worlds that survive the application of MP. The cognitive capacity of both MP-accessible models is reduced by the cognitive cost of applying MP, therefore $cp = (14, 5)$.

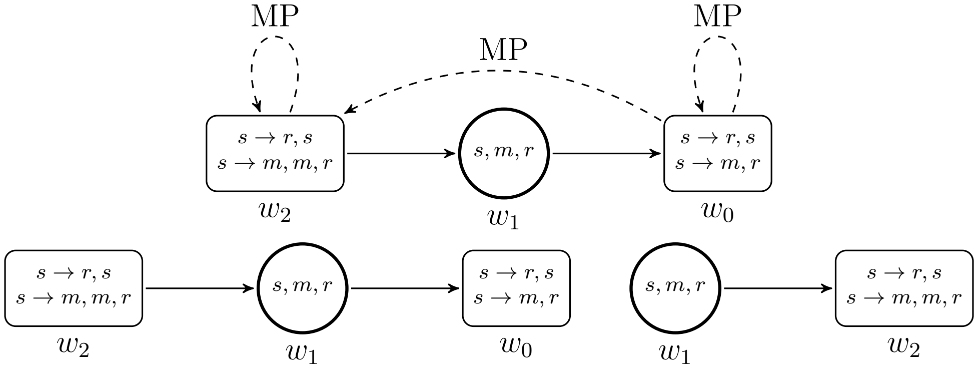


Figure 5.1: The first figure depicts the model M, with a MP-dashed arrow denoting that a world is an MP-expansion of another. Then, we obtain two potential transformations of the pointed model (M, w_1), i.e. two MP-accessible pointed models, based on the two ways the set of w_1's more (or equally) plausible worlds can change due to MP.

5.3.4 Truth clauses

We have explained how the original model changes after fast upgrades and slow applications of inference rules. Now come the truth conditions:

5.3.8. Definition (Dual process truth clauses). The following inductively define when a formula ϕ is *true at w in M* (notation: $M, w \models \phi$). For $w \in W^I$: $M, w \models \phi$ iff $\phi \in V_I(w)$. For $w \in W^P$:

$$\begin{array}{lll}
M, w \models p & \text{iff} & p \in V_P(w), \text{ where } p \in \Phi \\
M, w \models \neg\phi & \text{iff} & M, w \not\models \phi \\
M, w \models \phi \wedge \psi & \text{iff} & M, w \models \phi \text{ and } M, w \models \psi \\
M, w \models \Box\phi & \text{iff} & M, w' \models \phi \text{ for all } w' \in W \text{ such that } w \geq w' \\
M, w \models B\phi & \text{iff} & M, w' \models \phi \text{ for all } w' \in min(W) \\
M, w \models [\psi \Uparrow]\phi & \text{iff} & M^{\Uparrow\psi}, w \models \phi \\
M, w \models \langle\rho\rangle\phi & \text{iff} & M', w \models \phi \text{ for some } (M', w) \text{ which is } \rho\text{-accessible from } (M, w)
\end{array}$$

Logical validity is defined in terms of possible worlds only: a formula is valid in a model iff it is true at every possible world.

In accordance with what the dual process theories prescribe, our System 1 actions affect what is (defeasibly) known or believed without checking whether there is valid reasoning supporting the piece of information. Notice how this fits phenomena that are often seen as manifestations of System 1 being in charge, e.g. witnessed by experiments on the belief bias. Recall that these demonstrate that subjects are reluctant to believe "unbelievable" – given their prior conceptions – statements even when they logically follow from a set of premises. They also tend to believe "believable" conclusions, even though the underlying reasoning is problematic, due to the influence of pre-existing impressions and biases. These are hardwired in the model's plausibility ordering, while the fast upgrades integrate information based on them, thus forming the agent's epistemic or doxastic state without engaging in the effortful task of assessing what is valid. This falls under the responsibility of System 2; if the agent comes to know or believe something new following an action of System 2, this must follow logically from what is already known or believed.

Now we can develop our initial example into:

5.3.9. Example. Recall the scenario of Example 5.3.7. It is now easy to see that, based on our semantics, $\neg\Box m$, $\neg Bm$, $\Box s$, Bs, $\Box r$, Br are all valid; initially, Jill does not know, nor believes that m, despite knowing and believing that s. In addition, $\langle MP\rangle\Box m$, $\langle MP\rangle Bm$, $\langle MP\rangle\neg\Box m$, $\langle MP\rangle\neg Bm$ are all valid. That is, there is some application of MP that provides Jill with knowledge and belief of m (because she inferred it from $s \to m$ and s) and another application of MP that does not provide her with any new information (because she merely used $s \to r$ and s as premises, which only comes as a confirmation of her already held belief and knowledge of r).

The example shows how different applications of a rule, captured as different choices of expansions, may lead to different developments of the agent's knowledge and beliefs. Notice that the reading of $\langle\rho\rangle\phi$ is existential: it asks that there be *some* application of ρ leading to ϕ. Different choices allow both informative and uninformative applications by a competent agent with sufficient resources. The dual $[\rho]\phi := \neg\langle\rho\rangle\neg\phi$ is read as "after all applications of ρ, ϕ is true". This is satisfied whenever all ρ-accessible pointed models validate ϕ. Using $[\rho]$-operators, the modeller may express the overall effect of the rule to the agent's reasoning.

The previous example illustrated a simple case where slow thinking is affordable and the reasoning step of *Modus Ponens* is performed. In the next example, we model a *failure* to apply *Conjunction Introduction* (CI), following an application of *Double Negation Elimination* (DNE) and *Modus Ponens* (MP). This is illustrative of a depletion of resources that would halt the reasoning processes of System 2 and make the agent fall back to System 1. It corresponds to a series of examples offered by (Kahneman, 2011, Chapter 2): whenever the mental effort that System 2 requires wears the agent out completely, then she retreats to default System 1 activity.

5.3.10. Example.

- Take DPPM $M = \langle W^P, W^I, ord, V_P, V_I, cp\rangle$ parameterized by $R = \{DNE, MP, CI\}$, $Res = \{time, memory\}$ with $c(MP) = c(CI) = (1,2), c(DNE) = (3,1)$. In addition, take $cp = (4,7)$ and suppose that for $w \in W^P$: $M, w \models \Box\neg\neg\phi \wedge \Box(\phi \rightarrow \psi)$.

- Then, $M, u \models \neg\neg\phi$ and $M, u \models \phi \rightarrow \psi$ for all u such that $w \geq u$. Because of Succession, there is a model M' with $cp' = cp - c(DNE) = (1,6)$ such that $M', w \models \Box\phi$.

- Following the same procedure for MP, we get a model M'' with $cp'' = cp' - c(MP) = (1,6) - (1,2) = (0,4)$ such that $M'', w \models \Box\psi$.

- But then there cannot be any CI-accessible pointed model as the step is not affordable (compare $c(CI)$ and cp'').

- So finally, $M'', w \not\models \langle CI\rangle\Box(\phi \wedge \psi)$, therefore $M'', w \models \neg\langle CI\rangle\Box(\phi \wedge \psi)$. But this means that $M', w \models \langle MP\rangle\neg\langle CI\rangle\Box(\phi \wedge \psi)$.

- In turn, $M, w \models \langle DNE\rangle\langle MP\rangle\neg\langle CI\rangle\Box(\phi \wedge \psi)$.

- As a result, indeed $M, w \not\models [DNE][MP]\langle CI\rangle\Box(\phi \wedge \psi)$.

Before moving on to applications of the model, we introduce Theorem 5.3.11 and Theorem 5.3.12. These cast light on reasoning processes involving both inference rules used by System 2, *provided that they are affordable*, and fast upgrades by System 1. They can be generalized for more upgrades, applications of rules, and thus number of premises. Theorem 5.3.12 also exemplifies the order-sensitivity of a reasoning process that is orchestrated by both systems.

5.3.11. Theorem (Reasoning from rules).

> *If ψ logically follows from $\{\phi_1, \dots, \phi_k\}$ by applying the rules $\rho_1, \dots, \rho_n \in R$ then $\langle \ddagger \rangle^{m_i} \Box \phi_i$ (where $1 \leq i \leq k$ and $\langle \ddagger \rangle^{m_i}$ is a sequence of m_i-many inference rules available to the agent) implies $\langle \ddagger \rangle^{m_1} \dots \langle \ddagger \rangle^{m_k} \langle \rho_1 \rangle \dots \langle \rho_n \rangle \Box \psi$.*

Proof:
Take arbitrary DPPM M and world $w \in W^P$ of the model. Suppose $M, w \models \langle \ddagger \rangle^{m_i} \Box \phi_i$, for $1 \leq i \leq k$. For each ϕ_i, there is a model M^i such that $M^i, w \models \Box \phi_i$ which has $W^i = W \setminus \{u \in P_{\geq}(w) \mid u \notin P^i(w)\}$ where

- $P^i(w) = \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^i \mid v \in P_{\geq}(w)\})$
- $v^i = \{u \mid V^*(v) \vdash_{\langle \ddagger \rangle^{m_i}} V^*(u)\}$
- $V^*(v) \vdash_{\langle \ddagger \rangle^{m_i}} V^*(u)$ denotes that u is a $\rho_{i1}, \dots, \rho_{im_i}$-expansion of v for $\rho_{i1}, \dots, \rho_{im_i}$ composing the sequence $\langle \ddagger \rangle^{m_i}$.

This means that for all $u \in W^i$ such that $w \geq u$, $M^i, u \models \phi_i$. Due to Succession, there is some model M^* such that for all $u \in P^*(w)$: $M^*, u \models \phi_i$, for all $1 \leq i \leq k$, where $W^* = W \setminus \{u \in P_{\geq}(w) \mid u \notin P^*(w)\}$ with:

- $P^*(w) = \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^* \mid v \in P_{\geq}(w)\})$
- $v^* = \{u \mid V^*(v) \vdash_{\langle \ddagger \rangle^{m_i} \dots \langle \ddagger \rangle^{m_k}} V^*(u)\}$
- $V^*(v) \vdash_{\langle \ddagger \rangle^{m_1} \dots \langle \ddagger \rangle^{m_k}} V^*(u)$ denotes that u is a $\rho_{11}, \dots, \rho_{1m_1}, \dots, \rho_{k1}, \dots, \rho_{km_k}$-expansion of v.

Next, from the fact that ψ logically follows from $\{\phi_1, \dots, \phi_k\}$ through applying $\rho_1, \dots, \rho_n \in R$, and Succession, we get that there is a model $M^{\circledast}$ such that for all $u \in P^{\circledast}(w)$: $M^{\circledast}, u \models \psi$, which has $W^{\circledast} = W \setminus \{u \in P_{\geq}(w) \mid u \notin P^{\circledast}(w)\}$ where

- $P^{\circledast}(w) = \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^{\circledast} \mid v \in P_{\geq}(w)\})$
- $v^{\circledast} = \{u \mid V^*(v) \vdash_{\langle \ddagger \rangle^{m_i} \dots \langle \ddagger \rangle^{m_k}, \rho_1, \dots, \rho_n} V^*(u)\}$
- $V^*(v) \vdash_{\langle \ddagger \rangle^{m_1} \dots \langle \ddagger \rangle^{m_k}, \rho_1, \dots, \rho_n} V^*(u)$ denotes that u is a $\rho_{11}, \dots, \rho_{1m_1}, \dots, \rho_{k1}, \dots, \rho_{km_k}, \rho_1, \dots, \rho_n$-expansion of v.

But then $M^{\circledast}, w \models \Box \psi$, and overall $M, w \models \langle \ddagger \rangle^{m_1} \dots \langle \ddagger \rangle^{m_k} \langle \rho_1 \rangle \dots \langle \rho_n \rangle \Box \psi$. □

5.3.12. Theorem (Reasoning from upgrades and rules).

> *If χ logically follows from $\{\phi_1, \phi_2\}$ by applying the rules $\rho_1, \dots, \rho_n \in R$, then $[\psi \Uparrow](\Box \phi_1 \land \langle \ddagger \rangle^m \Box \phi_2)$ implies $[\psi \Uparrow] \langle \ddagger \rangle^m \langle \rho_1 \rangle \dots \langle \rho_n \rangle \Box \chi$.*

Proof:
Take arbitrary DPPM M and world $w \in W^P$ of the model. Suppose $M, w \models [\psi \Uparrow] (\Box\phi_1 \wedge \langle\ddagger\rangle^m \Box\phi_2)$. This amounts to $M^{\psi\Uparrow}, w \models (\Box\phi_1 \wedge \langle\ddagger\rangle^m \Box\phi_2)$, i.e. $M^{\psi\Uparrow}, u \models \phi_1$ for all $u \in P^{\psi\Uparrow}(w)$ (1) and there is a model M^* such that $M^*, w \models \Box\phi_2$ which has $W^* = W \setminus \{u \in P^{\psi\Uparrow}(w) \mid u \notin P^*(w)\}$ where

- $P^*(w) = \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^* \mid v \in P^{\psi\Uparrow}(w)\})$

- $v^* = \{u \mid V^*(v) \vdash_{\langle\ddagger\rangle^m} V^*(u)\}$

Then, $M^*, u \models \phi_2$, for all $u \in P^*(w)$. Due to Succession and (1), $M^*, u \models \phi_1$, for all $u \in P^*(w)$. Due to χ following from $\{\phi_1, \phi_2\}$ and Succession, there is a model $M^{\circledast}$ such that $M^{\circledast}, u \models \chi$, for all $u \in P^{\circledast}(w)$ where

- $P^{\circledast}(w) = \mathsf{C}$ where C is some choice in $\mathcal{C}(\{v^{\circledast} \mid v \in P^{\psi\Uparrow}(w)\})$

- $v^{\circledast} = \{u \mid V^*(v) \vdash_{\langle\ddagger\rangle^m, \rho_1, \ldots, \rho_n} V^*(u)\}$

But then $M^{\circledast}, w \models \Box\chi$, and overall $M, w \models [\psi \Uparrow]\langle\ddagger\rangle^m \langle\rho_1\rangle \ldots \langle\rho_n\rangle \Box\chi$. □

By making the semantic interpretations of propositional attitudes quantify over impossible worlds, it is guaranteed that some consequences of the agent's knowledge or beliefs are not known or believed: logical omniscience is thus avoided. Unlike other approaches though, the problem is escaped in a balanced manner, committed to the idea that people's mental states are formed by a variety of processes and their evolution depends on the availability of resources.

In view of considerations coming in Section 5.5 notice that one can read our models as *normative*, but *realistic*: an agent *ought* to choose and apply slow thinking rules to the extent that she *can* do it, given the cognitive resources at hand, and until these are depleted. Before we get there, in the next section, we put the framework to work.

5.4 Case studies

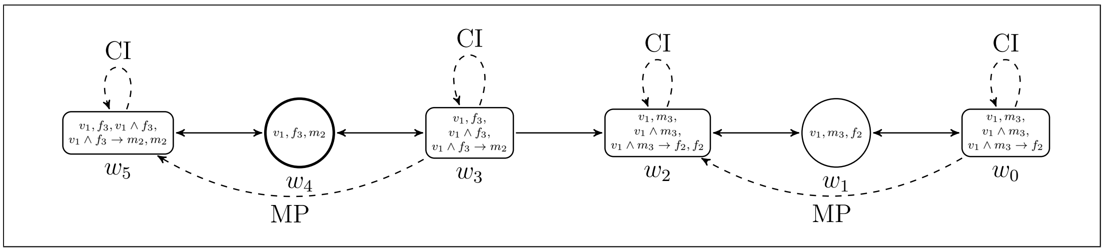
We now apply the dual process modelling to three case studies. In order to illustrate the workings of the two systems, we will assume that the applications of rules are all affordable to the involved agents.

Interaction between System 1 and System 2 (or, stereotypes gone wrong): System 1 provides its – sometimes incorrect – impressions to System 2. These impressions exemplify biases that are often attributed to our experience, the so-called *familiarity heuristics*. System 2 can then unpack their logical consequences. It is not uncommon for System 2 to eventually override System 1.

To demonstrate this, we introduce and analyze a variant of the *restaurant scenario*. Its original version goes as follows: “You are in a restaurant with your parents, and you have ordered three dishes: Fish, Meat, and Vegetarian. Now a new waiter comes back from the kitchen with three dishes. What will happen?” (van Benthem, 2008a, p.73). However, the evolution of the scenario might well be influenced by what is *familiar* to the waiter. Consider the following:

> *Jack (customer 1) and Jill (customer 2) have entered a restaurant. They are joined by John (customer 3) shortly after. Waiter A takes their order, which includes three dishes: Vegan, Meat and Fish. Waiter B is supposed to serve them. Waiter B is acquainted with Jack: he knows that Jack is a passionate animal rights activist, often arguing against the consumption of any animal product. He has not met Jill but he has the impression that she is pretty close to Jack and implicitly assumes that she shares his opinion and lifestyle. On the other hand, John is a regular customer: almost every time he orders the same meat-based dish. As the meals are prepared, Waiter B has an intuitive, yet incomplete, idea on their distribution. System 1 is at work. Influenced by his stereotypes and experience, he thinks that Jack, the vegan activist, will definitely get the vegan dish and that John will take the meat dish, as usual. For someone* carefully *and* consciously *reading the story, this would entail that Jill ordered fish. Not for waiter B, though. Busy as he is, and due to Jill's closeness to Jack, he has trouble inferring this conclusion. He is also willing to consider, albeit reluctantly, that John gets fish. Finally, cases where Jack orders meat or fish are ruled out by the waiter.*

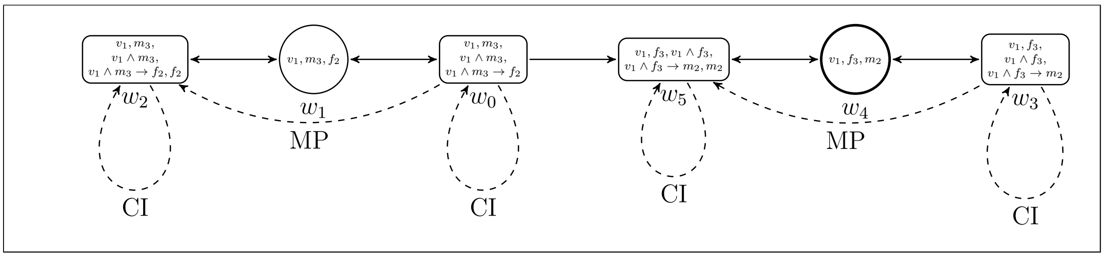
Let v_i, m_i, f_i ($i = 1, 2, 3$) denote the atoms expressing which dish goes to which customer. Let R be the set of rules containing *Conjunction Introduction* (CI) and *Modus Ponens* (MP) such that the rules are affordable to the agent (i.e. the capacity of the agent exceeds their costs). The following figure depicts the DPPM for waiter B.[10] Notice that the plausibility ordering is indicated by the solid arrows. According to our semantics, both Bv_1 and Bm_3 are valid.


CI
CI
CI
CI
$v_1, f_3, v_1 \wedge f_3,$ $v_1 \wedge f_3 \rightarrow m_2, m_2$
v_1, f_3, m_2
$v_1, f_3,$ $v_1 \wedge f_3,$ $v_1 \wedge f_3 \rightarrow m_2$
$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2, f_2$
v_1, m_3, f_2
$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2$
w_5
w_4
w_3
w_2
w_1
w_0
MP
MP

[10] Here, we took CI arrows to be reflexive and wrote down only the conjunctions obtained between atoms to increase the readability of the figure.

> *"John got fish this time!", says Waiter A. Waiter B overhears the comment and instantly incorporates this new piece of information.*

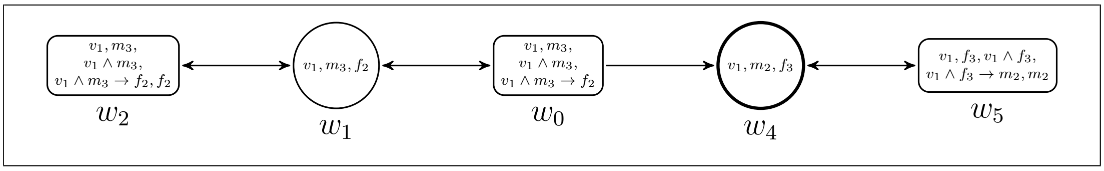
The model after the upgrade with f_3 is depicted below and is the outcome of combining the waiter's already held opinions and incoming information. System 1 deals with what is believed, on the basis of incoming information and biases generated by familiarity and experience, without investigating what follows logically.



$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2, f_2$
w_2
CI
v_1, m_3, f_2
w_1
MP
$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2$
w_0
CI
$v_1, f_3, v_1 \wedge f_3,$ $v_1 \wedge f_3 \rightarrow m_2, m_2$
w_5
CI
v_1, f_3, m_2
w_4
MP
$v_1, f_3,$ $v_1 \wedge f_3,$ $v_1 \wedge f_3 \rightarrow m_2$
w_3
CI

> *As Waiter B prepares to serve our three customers, he takes a moment to figure out what Jill actually ordered, contrary to what he would have expected. In particular, he realizes that Jill's relationship with Jack interfered with his beliefs. Instead, he should infer what follows from what he already believes, i.e. that Jill got the meat-based dish after all! This is due to a conscious procedure of System 2.*

Following an application of CI and MP, in that order, it is easy to verify that overall $[f_3 \Uparrow]\langle CI\rangle\langle MP\rangle Bm_2$ is valid. For example, the final pointed DPPM based on w_4 has worlds eliminated as epistemic possibilities by slow thinking: it exemplifies how System 2 took over System 1.

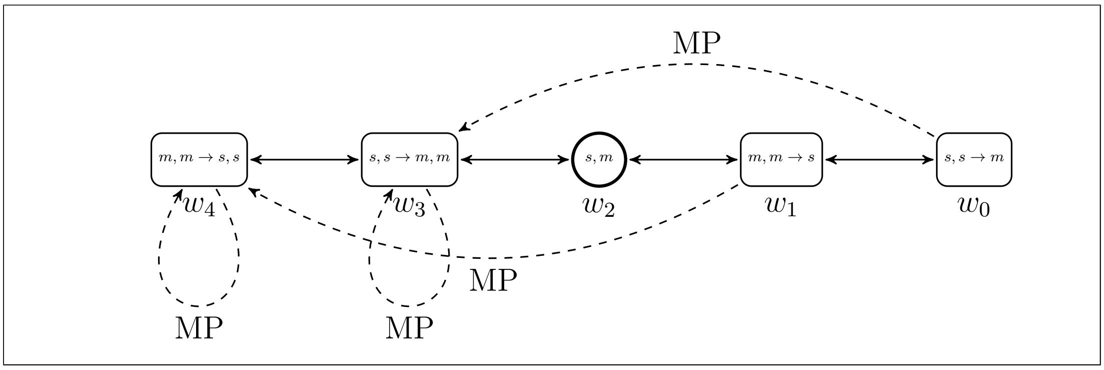


$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2, f_2$
w_2
v_1, m_3, f_2
w_1
$v_1, m_3,$ $v_1 \wedge m_3,$ $v_1 \wedge m_3 \rightarrow f_2$
w_0
v_1, m_2, f_3
w_4
$v_1, f_3, v_1 \wedge f_3,$ $v_1 \wedge f_3 \rightarrow m_2, m_2$
w_5

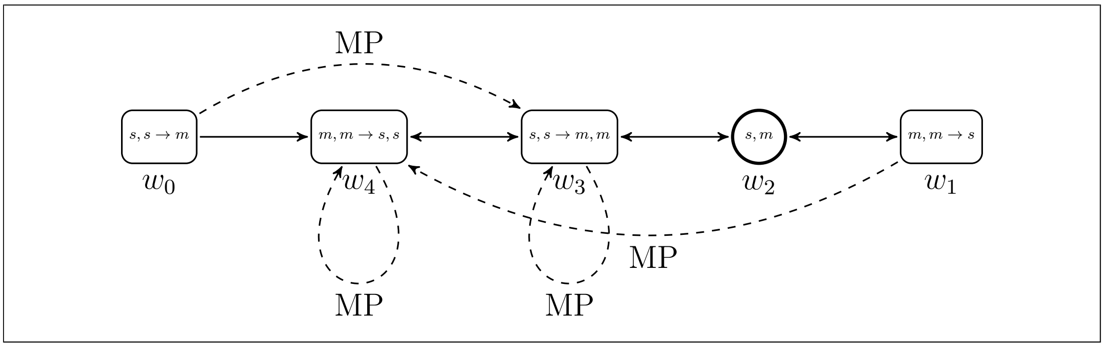
Framing effect: Decision-making is heavily influenced by the mode of presentation of options (Kahneman, 2011, Part 4). For instance, different responses are evoked whenever a question on the outcome of a surgery is presented in terms of survival or in terms of mortality. The statements "the survival rate within one month of surgery is 90%" and "mortality within one month of surgery is 10%" are equivalent: they have the same truth conditions. But under the first frame or mode of presentation, the situation seems somewhat more reassuring.

The framing effect poses a challenge for standard (D)EL. Propositional attitudes towards logically equivalent statements are the same under possible-worlds semantics, due to the closure properties of possible worlds. Also, according to the AGM approach to belief revision (Alchourrón et al., 1985), the beliefs of an agent are represented by a set of sentences in a formal language. This set is taken as closed under logical consequence. Therefore, if two sentences p and q are logically equivalent, then believing the one amounts to believing the other, and revising one's beliefs after being informed that p gives the same outcome as revising them after being informed that q. This too disregards the influence of framing on Humans, as opposed to Logons.

We will now show that framing can fit into our logical framework.[11] Let s and m denote the two statements discussed earlier (survival/mortality rate). Since the statements are equivalent, $s \leftrightarrow m$ is valid in our dual process semantics. Moreover, the agent has MP available and affordable, as in the previous case study. The initial model for our agent is given below and it can be verified that $\neg Bm$ and $\neg Bs$.

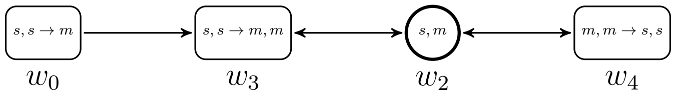


Following an upgrade with m, based on something the agent read in her social media, we obtain the model below. Therefore $[m \Uparrow]Bm$. As a result of framing, the agent has upgraded with m and believed in it, without simultaneously believing in s.



Again, some slow reasoning performed by System 2 will help the agent overcome framing: by performing an inference using Modus Ponens, the agent can come to believe that s too.

[11] In the context of this attempt, we model framing in an epistemic-doxastic setting but our tools can be aligned with dynamic preference logics (van Benthem, 2011; Liu, 2008, 2011) and hence framing effects on an agent's preferences (instead of beliefs) can be accounted for.



Anchoring effect: The *anchoring effect* (Tversky and Kahneman, 1975) is a cognitive bias that makes human reasoners rely heavily on the first piece of information they receive: this piece works as an "anchor", and even if it is clearly arbitrary and irrelevant, it can over-influence the formation of subsequent beliefs. For example, suppose that an agent is interested in a new edition of a high-end smartphone but has not made up her mind on whether to purchase it. The agent considers three options:

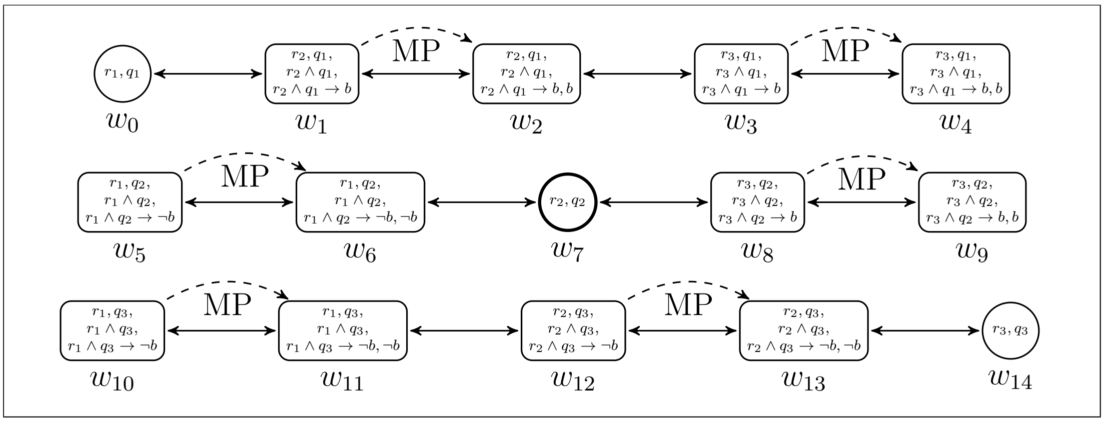
- r_1: the new edition falls in the price range [1000-1100).
- r_2: the new edition falls in the price range [1100-1200).
- r_3: the new edition falls in the price range [1200-1300).

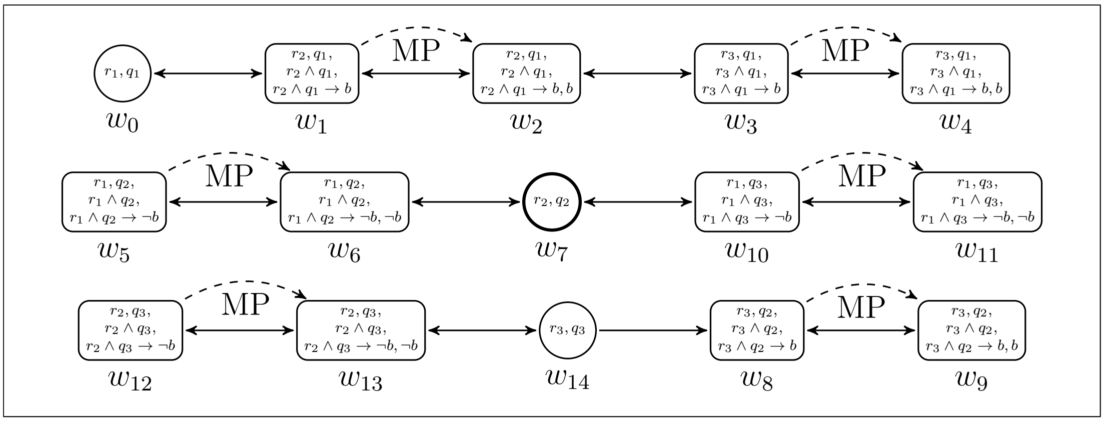
Suppose that the agent visits a store. She entertains the following options:

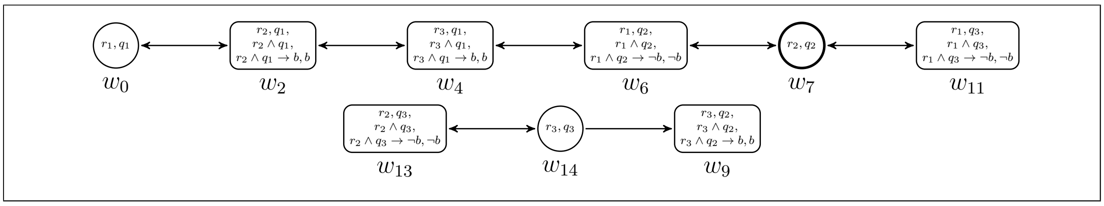
- q_1: the store's offer is in the price range [1000-1100).
- q_2: the store's offer is in the price range [1100-1200).
- q_3: the store's offer is in the price range [1200-1300).

In the store, there is a tag indicating that the original price of the desired item is, supposedly, 1200, but the store offers it for 1100. As a result, the agent performs a fast System 1 upgrade with the formula $[(r_3 \wedge q_2) \Uparrow]$. The value 1200 works as the anchor, because it is indicated by the store's tag as the market price of the new phone. As a result, the formula $[(r_3 \wedge q_2) \Uparrow]B(r_3 \wedge q_2)$ is valid.

Next, System 2 takes over and performs a reasoning step that allows her to believe that she saves a certain amount of money, which makes the bargain good (denote "good bargain" by b; also note that whenever $r_i \wedge q_i$, we consider the difference of prices negligible and thus not substantial enough to make the agent consider it a bargain). Therefore, we obtain a new valid formula: $[(r_3 \wedge q_2) \Uparrow]\langle MP \rangle Bb$. Based on that belief, she eventually acts accordingly and buys the smartphone. If there was no indication of an original market price of the smartphone or if the anchor was an initial value that the agent had set (i.e. deciding that only prices in the range [1000-1100) are acceptable), the evolution of the scenario would have been different and no purchase would have been made. Below, there is a depiction of the initial model, succeeded by the model following the anchoring upgrade, and one final model after the (affordable) MP-application.


r_1, q_1
w_0
$r_2, q_1, r_2 \wedge q_1, r_2 \wedge q_1 \to b$
w_1
MP
$r_2, q_1, r_2 \wedge q_1, r_2 \wedge q_1 \to b, b$
w_2
$r_3, q_1, r_3 \wedge q_1, r_3 \wedge q_1 \to b$
w_3
MP
$r_3, q_1, r_3 \wedge q_1, r_3 \wedge q_1 \to b, b$
w_4
$r_1, q_2, r_1 \wedge q_2, r_1 \wedge q_2 \to \neg b$
w_5
MP
$r_1, q_2, r_1 \wedge q_2, r_1 \wedge q_2 \to \neg b, \neg b$
w_6
r_2, q_2
w_7
$r_3, q_2, r_3 \wedge q_2, r_3 \wedge q_2 \to b$
w_8
MP
$r_3, q_2, r_3 \wedge q_2, r_3 \wedge q_2 \to b, b$
w_9
$r_1, q_3, r_1 \wedge q_3, r_1 \wedge q_3 \to \neg b$
w_{10}
MP
$r_1, q_3, r_1 \wedge q_3, r_1 \wedge q_3 \to \neg b, \neg b$
w_{11}
$r_2, q_3, r_2 \wedge q_3, r_2 \wedge q_3 \to \neg b$
w_{12}
MP
$r_2, q_3, r_2 \wedge q_3, r_2 \wedge q_3 \to \neg b, \neg b$
w_{13}
r_3, q_3
w_{14}


r_1, q_1
w_0
$r_2, q_1, r_2 \wedge q_1, r_2 \wedge q_1 \to b$
w_1
MP
$r_2, q_1, r_2 \wedge q_1, r_2 \wedge q_1 \to b, b$
w_2
$r_3, q_1, r_3 \wedge q_1, r_3 \wedge q_1 \to b$
w_3
MP
$r_3, q_1, r_3 \wedge q_1, r_3 \wedge q_1 \to b, b$
w_4
$r_1, q_2, r_1 \wedge q_2, r_1 \wedge q_2 \to \neg b$
w_5
MP
$r_1, q_2, r_1 \wedge q_2, r_1 \wedge q_2 \to \neg b, \neg b$
w_6
r_2, q_2
w_7
$r_1, q_3, r_1 \wedge q_3, r_1 \wedge q_3 \to \neg b$
w_{10}
MP
$r_1, q_3, r_1 \wedge q_3, r_1 \wedge q_3 \to \neg b, \neg b$
w_{11}
$r_2, q_3, r_2 \wedge q_3, r_2 \wedge q_3 \to \neg b$
w_{12}
MP
$r_2, q_3, r_2 \wedge q_3, r_2 \wedge q_3 \to \neg b, \neg b$
w_{13}
r_3, q_3
w_{14}
$r_3, q_2, r_3 \wedge q_2, r_3 \wedge q_2 \to b$
w_8
MP
$r_3, q_2, r_3 \wedge q_2, r_3 \wedge q_2 \to b, b$
w_9


r_1, q_1
w_0
$r_2, q_1, r_2 \wedge q_1, r_2 \wedge q_1 \to b, b$
w_2
$r_3, q_1, r_3 \wedge q_1, r_3 \wedge q_1 \to b, b$
w_4
$r_1, q_2, r_1 \wedge q_2, r_1 \wedge q_2 \to \neg b, \neg b$
w_6
r_2, q_2
w_7
$r_1, q_3, r_1 \wedge q_3, r_1 \wedge q_3 \to \neg b, \neg b$
w_{11}
$r_2, q_3, r_2 \wedge q_3, r_2 \wedge q_3 \to \neg b, \neg b$
w_{13}
r_3, q_3
w_{14}
$r_3, q_2, r_3 \wedge q_2, r_3 \wedge q_2 \to b, b$
w_9

5.5 Coda: "Ought Implies Can"

We conclude with a general philosophical issue: are our models merely descriptive of some of the cognitive workings of human agents, or rather normative? In the latter case, how so, since they aim to avoid the idealization of agents as logically omniscient?

One may take the logical approach proposed above as roughly standing to static (**S5**) epistemic logic and AGM belief revision theory as Kahneman and Tversky (1979)'s prospect theory of rational choice stands to expected utility theory. Just like prospect theory, our logic of fast and slow thinking is more complex than its mainstream counterpart: it adds operators and parameters to the standard EL framework, in order to provide a more realistic account of reasoning by human agents. Complexity is generally taken as a theoretical cost, to be justified by a gain in explanatory and predictive power. Here we have an unavoidable trade-off. *Any* framework for EL needs to strike a balance between two desiderata. The pull towards simplicity and idealization leads in the direction of Logons. The pull towards modelling realistic Humans can easily lead to conceptually gerrymandered frameworks, or to logics that are too weak to be of serious interest. Take Jill, who knows that $\phi \wedge \psi$. What epistemic facts follow? She may fail to unpack her knowledge, so she need not know that ψ. She may also not know that χ, although χ turns out to be logically equivalent to the conjunction of ϕ and ψ.

The trade-off between simplification and realism overlaps that between description and prescription, evident in many disciplines. Prospect theory was justified as a descriptive theory of rational decision, in opposition to the normative status of classical expected utility theory. We have a more nuanced stance with respect to the logic proposed above. We aim at a *normative* logical theory; but, one whose rational "ought", unlike the "ought" of static epistemic logic, implies "can".

To unpack, let's reflect back to Chapter 2. The mainstream EL approach has been defended on basis of its normative status, analogies to models used in natural sciences, etc. The Standard Rationality Thesis, which corresponds to agents of the mainstream formalization, has also been defended in terms of the performance/competence distinction, etc. We have, however, explained why these arguments are unsatisfactory and we proposed an alternative rationality picture that takes into account empirically indicated limitations of real reasoners, the nuances of their mental states and of the processes underlying their explicit reasoning.

We have argued that this picture has implications for EL. The choice between models that are *either* merely descriptive, *or* representing omniscient agents is a false dilemma. Whereas the mainstream logical "ought" fails to imply "can", one may be interested in investigating an "ought" that does, without requiring a "brain the size of a blimp" (Stich, 1990, p.26). Logons know or believe the infinitely many logical consequences of what they know or believe. But the available resources of human agents are not infinite (Cherniak, 1986), and "become infinite" is a strange thing to ask of a finite mind.

In our approaches, empirical evidence can contribute in picking the appropriate normative model. Limitations in terms of time, memory, attention, etc., are important in adjusting the rationality standard expected from the agent's deductive reasoning. This was first addressed in Chapter 3. The variety of propositional attitudes and the processes underlying deduction were further addressed in this chapter and Chapter 4. Empirical data can be therefore utilized in constructing the right normative model, e.g. by filling in the right parameters. So we put forth our logical systems as better normative models, delivering a can-implying "ought". A finite and fallible, but rational, agent ought to reason to the extent that, *ceteris paribus*, her limited resources allow.[12] No more can be asked without violating that implication, but also no less.

5.6 Conclusions

To sum up, we have built an alternative logical system that avoids the problem of logical omniscience, which has plagued standard EL. Most importantly, it does so by taking on board a popular line of research in psychology of reasoning: dual process theories. Our system includes two different kinds of dynamic operators, one responsible for the fast and effortless integration of information and one accounting for the slow and costly steps of deductive reasoning. In order to accommodate the respective actions, tools from DEL were combined with impossible-worlds semantics. We demonstrated that this framework successfully captures desirable properties of reasoning processes composed by both systems. In particular, we showed that phenomena that have been studied in the literature of various disciplines can be now formally treated in logical terms. Our exposition was finally furnished with a philosophical discussion on the contribution of the attempts of Part II, and more specifically, on their normative nature.

The model deals only with a fragment of the activities undertaken by the two systems. Apart from adding probabilistic reasoning for a more elaborate modelling of System 2, other directions of further work can also be envisaged. First, the policy of upgrading with incoming information need not be unique. More conservative System 1 actions can be modelled, sensitive to the reliability of the source (van Benthem, 2011). Second, one may combine this work with the distinction between implicit acts of observation ("bare observation") and explicit acts of observation ("conscious realization") (van Benthem, 2008c,b; Velázquez-Quesada, 2009). This distinction can be accommodated by additional actions, of the sort introduced in Section 4.4. These will be representative of the two systems: the former kind is effortless and corresponds to System 1's fast processing of incoming information. The latter kind is resource-consuming and corresponds to System 2 activities. Third, one may model higher-order reasoning, accounting for how the agent thinks over her own reasoning processes, learns or forgets inference rules, that in turn affect her deductive inferences. The latter can be accommodated as in Section 4.4. Still, we have so far focused on how the agent expands her factual information without delving into higher-order reasoning. It seems that, in order to enrich the picture of reasoning run by System 2, we need to impose additional constraints on the model's structure and define suitable actions of rule-based and effortful higher-order reasoning. We will pursue these questions in Part III.

[12] The *ceteris paribus* parameter matters. What amount of cognitive resources ought to be allocated to reasoning tasks is heavily context-dependent: one should not be asked to deploy cognitive resources to perform logical deductions when this would make it dangerous to thoughtlessly cross a busy street.

Part III

Multi-agent Reasoning

Chapter 6

Formalizing false belief tasks

In this chapter, we address one dimension of reasoning in a multi-agent setting, which naturally highlights the importance of the notion of ToM.[1] In particular, we focus on the *formation* of beliefs about others' beliefs, shaped by observation, memory, and communication. Theory of Mind, and the paradigmatic tasks studying it, the False Belief Tasks (FBTs), are instrumental in this investigation (Section 2.3.3). Their formal study has become important, witness recent research on reasoning and information update by intelligent agents, and some proposals for its formal modelling have put forward settings based on EL. Still, due to its intrinsic idealizations, it is questionable whether EL can be used to model the higher-order cognition of "real" agents. This chapter proposes a framework for modelling mental state attributions that is more in-line with findings in cognitive science and applicable to FBTs. We introduce a temporal setting focusing on visibility, memory, and communication as factors underpinning mental attributions. We discuss some of its technical features, use it for modelling well-known FBTs, and argue on why it does justice to empirical observations.

6.1 Introduction

Theory of Mind is an important feature of how people function in social scenarios: someone who understands that others might have mental states different from hers and can reason about those states is much better suited to understand their behaviour, and thus act and react appropriately.[2]

[1] The chapter is based on Solaki and Velázquez-Quesada (2019, 2021).

[2] There has been a debate on how this understanding of others' mental states is achieved (see, e.g., Carruthers and Smith (1996)). Some argue that it is by acquiring a *theory* of commonsense psychology (*theory theory*); some others argue that it comes from a direct *simulation* of others' mental states (*simulation theory*). We will use the term ToM without endorsing any of these views, as such discussion falls outside the scope of this proposal.

Theory of Mind is slowly developed in the course of our lives (Wimmer and Perner, 1983; Wellman, 1991) and at a different speed for different types of persons (Baron-Cohen et al., 1985). The starting point consists of developing the ability to make *first-order* attributions (e.g. knowing/believing that *"Mary believes that the ball is in the bag"*), then moving on to attributions of *second-order* mental states (e.g. knowing/believing that *"Mary believes that John believes that the ball is in the closet"*), and so on. When testing the ToM of an individual, an extensively used experiment is the *Sally-Anne* False Belief Task.

6.1.1. Example (The *Sally-Anne* task). The following is adapted from (Baron-Cohen et al., 1985).

> *Sally and Anne are in a room in which there are a basket and a box. Sally is holding a marble. Then, after putting the marble into the basket, Sally leaves the room. While Sally is away, Anne transfers the marble to the box. Then Sally comes back.*

To pass the test, the subject should answer correctly the question *"where does Sally believe the marble is?"*. This requires for the subject to distinguish her own true belief (*"the marble is in the box"*) from Sally's *false* belief (*"the marble is in the basket"*). Experiments (e.g. Wimmer and Perner (1983)) have shown that, while children older than 4 years old tend to answer correctly, younger children (or children on the autism spectrum) tend to report their own belief, thus failing the test. (But see Setoh et al. (2016).)

Experiments such as the *Sally-Anne* task show that young children do not have basic ToM. But as we have seen in Section 2.3.3, adults have limitations too, confirming that ascribing beliefs to others is not automatic but it rather requires an effortful process that comes with its own cognitive costs.

In the enterprise of studying and understanding ToM, there has been a growing interest in the use of formal frameworks. Among logicians, a seemingly natural choice is EL, because it provides us with tools for representing not only the knowledge/beliefs agents have about ontic facts, but also the knowledge/beliefs they have about their own and other agents' knowledge/beliefs. However, using EL as the basic layer of a system for ToM has some drawbacks, explained in Section 2.2.2, e.g. the problem of logical omniscience, the unlimited (positive and negative) introspection and reasoning about others.

There is an even more fundamental reason why EL might not be well-suited for representing realistic higher-order attributions. Semantically, knowledge/beliefs are given by a universal quantification: ϕ is known/believed iff it holds in *all* the alternatives the agent considers epistemically/doxastically possible. But, for real agents, knowledge and belief involve more elaborate considerations, such as the existence of (a certain type of) justification or evidence, typically produced through reasoning, observation or communication. The "simple" universal quantification EL uses, works because its semantic model carries a large amount of information: not only the (maximally consistent) alternatives the agent considers possible, but also every other alternative *every other agent* considers possible.[3] In a few words, the EL semantic interpretation of (higher-order) knowledge/belief formulas is "simple" because the model representing the current stage contains already all the needed information. Real agents might not be able to have such a large structure "in their mind", and thus it is questionable whether traditional EL can properly represent the way real agents deal with mental attribution scenarios.

In light of these and other similar issues, one could even wonder whether it makes sense to use logical tools for dealing with results of empirical research on mental state attributions. It has been argued that psychological experiments and logic are essentially different, understanding the former as the study of empirical findings on the behaviour of real "fallible" agents, and the latter as a normative discipline studying what "rational" agents *should* do.[4] However, we have justified (Section 2.5.2) why bridging these two views is a worthwhile endeavour that also has promising applications (especially on reasoning and information update by intelligent agents). Indeed, while empirical research benefits from the use of logical tools to explain its discoveries and understand their consequences, logical frameworks become richer and more "useful" when they capture human limitations and prescribe behaviour attainable by real agents.

This proposal seeks a logical formalization that allows us to reason about ToM. More precisely, the goal is to model FBTs, as the paradigmatic ToM tasks, using a framework that is more in-line with findings in the cognitive science literature, and therefore resembles more closely the way humans make attributions of mental states.[5] Still, different frameworks have followed different intuitions and used different formal tools. In light of this diversity, it makes sense to ask the following question: which requirements should the system satisfy to be useful?

Bolander (2018) proposes two criteria for evaluating formalizations of FBTs: ***(i)*** *robustness* (being able to deal with as many FBTs as possible, with no strict limit on the order of belief attribution), and ***(ii)*** *faithfulness* (each action of the task should correspond to an action in the formalism in a natural way). Still, the current proposal wants to achieve something more. As mentioned before, we

[3] Frameworks for representing acts of private communication (Baltag et al., 1998) make this clear. Their additional structures, *action models*, have an "event" for each one of the different ways the agents might perceive the communication that is taking place. Then, the model that results from the communication contains roughly one copy of the original model for each one of these perspectives.

[4] *Anti-Psychologism* (e.g. Frege (1884)) has long been against attempts to reconcile the two (Pelletier et al., 2008).

[5] Existing approaches for formalizing FBTs (see Section 6.2.4, and also Bolander (2018)) have also looked for formal settings that can be used as the basis for the reasoning engine of autonomous agents. As discussed by Verbrugge (2009), these goals are not disjoint.

believe that the syntax and the semantics of a framework looking for a faithful representation of mental attributions should resemble closely the way humans carry out these tasks. If this is achieved, then the framework can be also used in another interesting way: it can help us identify reasons why people find these tasks increasingly difficult as their order grows, thus also helping us understand what goes wrong when people fail.

To that end, we aim at the converse direction to that of EL. Our structures will be simple, with the representation of a single stage encoding only very basic facts, resembling the "frugal" way real agents keep information stored. However, interpretations of mental state attributions will resemble the arduous and oftentimes strenuous process through which agents recall these facts and derive further information on their basis. Among the many factors that can play a role when an agent makes mental attributions, our framework will focus on three of them. The first is an explicit representation of what an agent can *observe*. The second is the agent's *memory* about previous stages, which she uses when the basic facts she currently has "at hand" are not enough. The third is *communication*, providing the agent with information she cannot obtain directly on her own. These aspects have been already considered separately by different logical frameworks for representing FBTs, and in general by logical frameworks dealing with information and its dynamics (Section 6.2.4). The current proposal will show how their combination plays a role in a wide variety of FBTs.

The text is organized as follows. Section 6.2 introduces the basic *temporal visibility* framework, which relies on two of the mentioned aspects: what an agent can observe at the current stage, and what she remembers about visibility in the past. The section presents the semantic model and the formal language, showing how they can be used to model well-known FBTs. After discussing some of the framework's technical aspects, the section closes by evaluating it in terms of the criteria presented above, relating its features with findings in the cognitive science literature, and comparing it with similar proposals.

Still, while the basic framework is enough for modelling some FBTs, it is not enough to model others, in particular those that rely not only on observations and memory, but also on communication. Section 6.3 thus extends the basic framework to capture this crucial aspect. The section is structured as the previous one: it starts by presenting the framework and using it for modelling FBTs involving communication, and then it discusses some of its technical and cognitive features. Section 6.4 closes this proposal, recapitulating the highlights and discussing lines for its further development.

6.2 Visibility in a temporal setting

We start off with a temporal setting addressing mental state attributions formed due to visibility of basic facts and the agents' memory thereof. In most mental attribution tasks, beliefs[6] are, at their lower (ontic) order, about the location of certain objects (e.g. the marble's location in the Sally-Anne Task). We do take objects as the main entities about which agents have mental attitudes; still, for the simplicity of representation, we will work with these objects' *colours*. For example, in the Sally-Anne Task, the marble being located in the basket will be represented by taking its colour to be *white*.

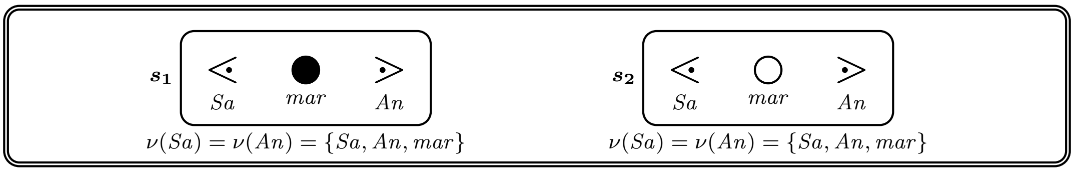
Throughout this chapter, let $A \neq \varnothing$ be the set of agents $(a, b, \ldots)$ and $O \neq \varnothing$ be the set of objects $(o, p, q, \ldots)$. For each $o \in O$, the set R_o contains the colours the object might have; define $R_O := \bigcup_{o \in O} R_o$. The model is a temporal structure, with each *state* fully described by both the colour of each object and the objects and agents each agent sees.

6.2.1. Definition (State and temporal visibility model). A *state* is a tuple $s = \langle \kappa, \nu \rangle$ where ***(i)*** $\kappa : O \to R_O$ is a *colouring* function, indicating the colour of each object,[7] and ***(ii)*** $\nu : A \to \wp(A \cup O)$ is a *visibility* function, indicating the *entities* (agents and objects) each agent sees.[8] The colouring and visibility functions of a state s will be denoted by $s.\kappa$ and $s.\nu$, respectively, and directly by κ and ν when no confusion arises.

A *temporal visibility model* (TVM) is a finite non-empty sequence of states $M = s_1 \cdots s_n$, with $\sharp_M := n$ its *cardinality* and $s_{last(M)} := s_{\sharp_M}$ its *last* state. We will write $s_i < s_j$ iff $i < j$.

Each state has components capturing the basic facts: the colour of the objects and, crucially, the visibility of the agents. Then, the model contains states, each intuitively corresponding to a different temporal stage. In this way, the model will account for the formation of beliefs grounded on visibility and memory.

6.2.2. Example. Take the Sally-Anne Task, with Sally (Sa), Anne (An) and the marble (mar). The story up to immediately before Sally leaves is represented by a two-state model M with ***(i)*** s_1 the initial state, where both agents see all agents and objects ($s_1.\nu(Sa) = s_1.\nu(An) = \{Sa, An, mar\}$) and the object is black ($s_1.\kappa(mar) = black$, with *black* indicating 'Sally's hands'), and ***(ii)*** s_2 the "next" state, where both agents still see everything, but now the object is white ($s_2.\kappa(mar) = white$, with *white* indicating 'the basket'). The model can be depicted as:


s_1
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$
s_2
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$

[6] Following the common parlance in the literature describing the tasks we later model, the term *belief* will be used for referring to an agent's mental state.

[7] Each object has a proper colour: $\kappa(o) \in R_o$ holds for all $o \in O$.

[8] Every agent can see herself: $a \in \nu(a)$ holds for all $a \in A$.

Representing actions A TVM M contains not only a state representing the current situation ($s_{\sharp_M}$) but also states indicating how the situation was in the past ($s_1, \ldots, s_{\sharp_M - 1}$). Still, situations also evolve forward: in the Sally-Anne Task, some further acts modify the colour of the object (Sally puts the marble into the basket) and some others modify the agents' visibility (Sally leaves the room). To allow the representation of these actions, one can provide operations that *extend* the current model with a state depicting the outcome of the given activity, thus also representing the way the situation *will be*. Here are their definitions.

6.2.3. Definition (Colour change). Let $M = s_1 \cdots s_n$ be a TVM. Take a set of objects $\{o_1, \ldots, o_k\} \subseteq O$, with $c_i \in R_{o_i}$ $(1 \leqslant i \leqslant k)$ a proper colour for each one.

The colour assignment $C = [o_1{:=}c_1, \ldots, o_k{:=}c_k]$ produces the model $M_C = s_1 \cdots s_n s_{new}$, which extends M with a state s_{new} whose components are defined, for every $o \in O$ and $a \in A$, as

$$s_{new}.\kappa(o) := \begin{cases} c_i & \text{if } o = o_i \in \{o_1, \ldots, o_k\} \\ s_{last(M)}.\kappa(o) & \text{otherwise} \end{cases}, \qquad s_{new}.\nu(a) := s_{last(M)}.\nu(a)$$

Thus, while visibility in s_{new} is exactly as in $s_{last(M)}$, colouring in s_{new} follows C for the objects the assignment mentions, remaining exactly as $s_{last}(M)$ for the rest. The requirement on each c_i guarantees that M_C is indeed a TVM.

6.2.4. Definition (Visibility change). Let $M = s_1 \cdots s_n$ be a TVM. Take a set of agents $\{a_1, \ldots, a_h\} \subseteq A$, with sets $X_i \subseteq A \cup O$ $(1 \leqslant i \leqslant h)$ satisfying $a_i \in X_i$.

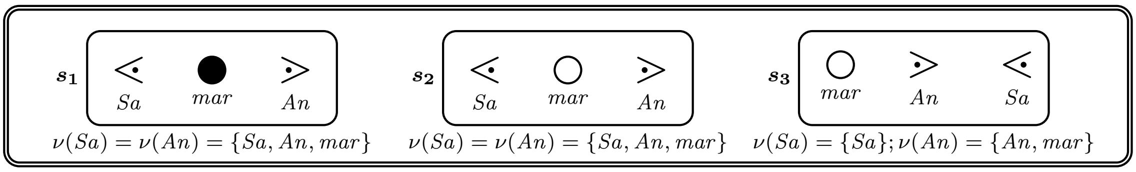
The visibility assignment $V = [a_1{\leftarrow}X_1, \ldots, a_h{\leftarrow}X_h]$ produces the model $M_V = s_1 \cdots s_n s_{new}$, which extends M with a state s_{new} whose components are defined, for every $o \in O$ and $a \in A$, as

$$s_{new}.\kappa(o) := s_{last(M)}.\kappa(o), \qquad s_{new}.\nu(a) := \begin{cases} X_i & \text{if } a = a_i \in \{a_1, \ldots, a_h\} \\ s_{last(M)}.\nu(a) & \text{otherwise} \end{cases}$$

Thus, while colouring in s_{new} is exactly as in $s_{last(M)}$, visibility in s_{new} takes the visibility of $s_{last}(M)$ for agents not mentioned by the assignment, following the assignment for the agents it mentions. The requirement on each X_i guarantees that M_V is indeed a TVM.

Let's see the operation through an example:

6.2.5. Example. Recall the Sally-Anne Task, with its first two stages represented by the TVM in Example 6.2.2. The story continues with Sally leaving the room; after this, she can see neither Anne nor the marble anymore, and Anne can only see the marble (and herself). This is represented by an operation extending the model with a new state (s_3) in which both Sa's and An's visibility have changed, yielding the model $M_{[Sa \leftarrow \{Sa\}, An \leftarrow \{An, mar\}]}$ below.


s_1
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$
s_2
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$
s_3
mar
An
Sa
$\nu(Sa) = \{Sa\}; \nu(An) = \{An, mar\}$

The operations describe a change in the current situation; in this sense, they are analogous to model transformations in DEL (van Ditmarsch et al., 2007; van Benthem, 2011). Still, there is an important difference. Typically, DEL models describe only the current situation, so model operations return a structure representing also a single situation (the "next" one). In contrast, while a TVM M describes how the situation is at the current stage (the state $s_{last(M)}$), it might also describe how the situation was in the past (the previous states, when they exist). Thus, while the operations add a state describing the situation the action produces, they also retain the states of the original model, hence keeping track of what happened before. This is instrumental for the intuition that belief attributions at a certain stage also depend on the agents' recollection of what happened earlier. In this sense, the TVM setting can be understood as a "dynamic temporal": an underlying temporal structure that can be *extended* by dynamic *model change* operations. Other logical frameworks using similar ideas include Yap (2011) (cf. Sack (2008); Renne et al. (2016)), which redefines the operation representing acts of (public and) private communication (Baltag et al., 1998) to preserve previous stages, and Baltag et al. (2018), whose models "remember" the initial epistemic situation.

A formal language The language $\mathcal{L}$, for describing TVMs, contains basic formulas expressing the (higher-order) beliefs agents have about the colour of an object, and it is closed under both Boolean operators and dynamic modalities (those describing what will be the case after an action takes place).

6.2.6. Definition (Language $\mathcal{L}$). Given sets A, O and $\{R_o\}_{o \in O}$ as before, formulas ϕ of the language $\mathcal{L}$ are given by:

$$\begin{array}{ll} \phi ::= \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c) \mid \neg\phi \mid \phi \wedge \phi \mid [\alpha]\,\phi & \text{for } m \geqslant 1,\ \{a_1, \ldots, a_m\} \subseteq A,\ o \in O,\ c \in R_o \\ \alpha ::= p_1{:=}c_1, \ldots, p_k{:=}c_k \mid b_1{\leftarrow}X_1, \ldots, b_h{\leftarrow}X_h & \text{for } k \geqslant 1,\ \{p_1, \ldots, p_k\} \subseteq O,\ c_i \in R_{p_i}, \\ & h \geqslant 1,\ \{b_1, \ldots, b_h\} \subseteq A,\ X_i \subseteq A \cup O,\ b_i \in X_i \end{array}$$

Formulas of the form $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$, called *mental attribution formulas*, are read as *"agent a_1 believes that ...that agent a_m believes that o has colour c"*. Other Boolean connectives ($\vee, \rightarrow, \leftrightarrow$) are defined in the standard way.

Formulas in $\mathcal{L}$ are evaluated in a TVM with respect its last state, the fullest representation of the scenario available up that point. Nevertheless, as the definition shows, the truth-value of formulas is influenced by earlier states.

6.2.7. Definition (Semantic interpretation). Let $M = s_1 \cdots s_n$ be a TVM. The following abbreviation will be useful.

- Take a mental attribution formula $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$. Its *visibility condition* on a state s, denoted by $\mathrm{vis}_\chi(s)$, lists the requirements for χ to be evaluated at s (agent a_1 can see agent a_2, …, agent a_{m-1} can see agent a_m, agent a_m can see object o). Its formal definition is

$$\mathrm{vis}_\chi(s) \quad \textit{iff}_{\,def} \quad a_2 \in s.\nu(a_1) \;\&\; \ldots \;\&\; a_m \in s.\nu(a_{m-1}) \;\&\; o \in s.\nu(a_m)$$

For evaluating $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$ at a TVM, the process starts on the model's last state, going "back in time" one step at the time, looking for a state satisfying χ's visibility condition. If such state s' is reached, χ's truth-value depends only on whether o has colour c at s'; otherwise, χ is false. Formally, and by using "⅋" for a natural-language disjunction (just as "&" stands for a natural-language conjunction), the satisfaction relation $\Vdash$ between a temporal visibility model $M = s_1 \cdots s_n$ and a mental attribution formula is given by

$$M \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c) \quad \textit{iff}_{\,def} \quad \bigparr_{i=0}^{n-1} \left(\begin{array}{c} \overbrace{\mathrm{vis}_{\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)}(s_{n-i})}^{\mathsf{vis}} \;\&\; \overbrace{s_{n-i}.\kappa(o) = c}^{\mathsf{col}} \\ \& \\ \underbrace{\mathop{\&}_{j=1}^{i} \text{ not } \mathrm{vis}_{\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)}(s_{n-(j-1)})}_{\mathsf{no-latter-vis}} \end{array} \right)$$

Thus, $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$ holds at M when there is a state (the quantification indicated by the main disjunction) in which the visibility condition is satisfied (the vis part), the object has the indicated colour (the col part), and there is no "more recent" state satisfying the visibility condition (the no−latter−vis part).

Boolean operators are interpreted as usual. For dynamic modalities,

$$M \Vdash [\alpha]\, \phi \qquad \textit{iff}_{\,def} \qquad M_{[\alpha]} \Vdash \phi$$

There are some points about the semantic evaluation that are worthwhile to emphasize. ***(i)*** The semantic interpretation of $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$ captures the discussed intuitive idea. On the one hand, if the visibility condition fails at every state, the formula is false (every disjunct fails in its vis part). On the other hand, if some states satisfy the visibility condition, let s_ℓ be the time-wise latest (i.e. $\ell = \max\{i \mid s_i \in M \text{ and } \mathrm{vis}_\chi(s_i)\}$); then, $M \Vdash \chi$ iff $s_\ell.\kappa(o) = c$. ***(ii)*** For the sake of simplicity, we assume that, when an agent a sees an agent b, and b sees an object o, then a in fact sees b *seeing* o, as it should be intuitively the case in order for a formula like $\mathrm{B}_a\,\mathrm{B}_b(o \lhd c)$ to be evaluated.[9] ***(iii)*** Still, the visibility of the agents is not "common knowledge": an agent can see without being seen (see Section 6.2.1). ***(iv)*** The term *belief* here does not have the strong EL reading; it is rather understood as *"truth according to the agent's current information about what has happened so far"*. In this sense, it follows a form of *default reasoning* (Reiter, 1980; Ben-David and Ben-Eliyahu-Zohary, 2000): the agent assumes that things remain the way she saw them last. ***(v)*** Given the self-visibility requirement on the visibility function, attributions to oneself boil down to the col part of the interpretation, thus giving any agent full positive introspection (this will be addressed in Chapter 7).

6.2.1 Modelling FBTs

In what follows, we provide formal representations of two well-known FBTs.

6.2.8. Example (First-order FBT: the *Sally-Anne* task). The task's full story (Example 6.1.1) can be represented within the TVM framework. ***(i)*** Sally and Anne are in a room, with Sally holding the marble (a model with only state s_1 of Example 6.2.2). ***(ii)*** Sally puts the marble into the basket (the full model in Example 6.2.2). ***(iii)*** Sally leaves the room (the model in Example 6.2.5). ***(iv)*** Anne transfers the marble to the box, with *green* indicating "the box" (the model in Figure 6.1). The task's last step, Sally coming back to the room, prepares the audience for the crucial question: *"where does Sally believe the marble is?"*. The action changes Sally's visibility (she can see Anne now), but it does not change the crucial fact that she cannot see the marble. Thus, it is not relevant for mental states attributions.

So, which are Anne's and Sally's higher-order beliefs at the end of the story? According to the framework, with M the model in Figure 6.1,

- $M \Vdash \mathrm{B}_{An}(mar \lhd green)$ because, at s_4, Anne sees the marble ($mar \in s_4.\nu(An)$) and the marble is indeed green ($s_4.\kappa(mar) = green$).
- $M \Vdash \mathrm{B}_{Sa}(mar \lhd white)$ because, although Sally cannot see mar now (at s_4), mar was white the last time she saw it (s_2).
- $M \Vdash \mathrm{B}_{Sa}\,\mathrm{B}_{An}(mar \lhd white)$ because the last time Sally saw Anne seeing mar (s_2), mar was white.
- $M \Vdash \mathrm{B}_{An}\,\mathrm{B}_{Sa}(mar \lhd white)$ because the last time Anne saw Sally seeing mar (s_2), mar was white.

[9] So, even though visibility is not transitive (agent a_1 might see agent a_2, and the latter might see an object o, but this does not imply that a_1 sees o), there is some form of "transitivity assumption": if a_1 sees a_2 and a_2 sees o, agent a_1 might not see o, but she sees a_2 seeing o. This assumption might be a problem for attributions under (semi-)private actions. The work of Gasquet et al. (2016) and Charrier et al. (2016) can be especially relevant in that respect.

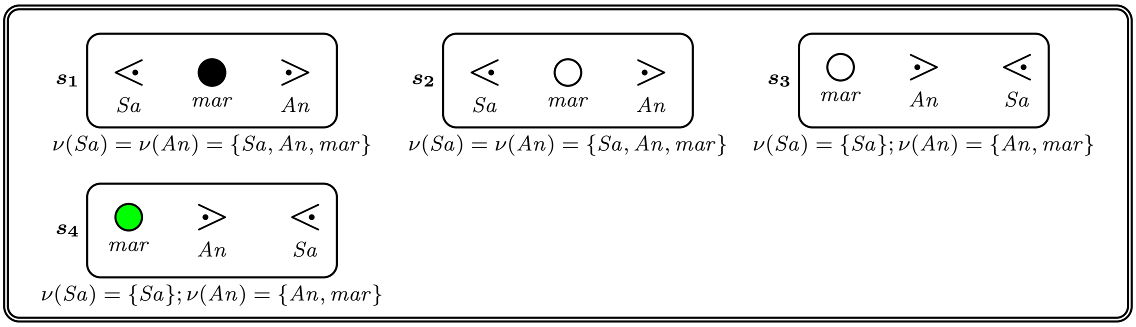

s_1
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$
s_2
Sa
mar
An
$\nu(Sa) = \nu(An) = \{Sa, An, mar\}$
s_3
mar
An
Sa
$\nu(Sa) = \{Sa\}; \nu(An) = \{An, mar\}$
s_4
mar
An
Sa
$\nu(Sa) = \{Sa\}; \nu(An) = \{An, mar\}$

Figure 6.1: TVM representation of the full Sally-Anne Task.

The full modelling of the Sally-Anne Task shows how the framework can be used to model attributions determined by the agent's visibility of both objects and other agents. The crucial role of memory is encoded in our semantic interpretation: in the last three formulas evaluated, the necessary information was not available at the last (most recent) state. As a result, the agents followed a backtracking process to s_2 and grounded their beliefs on facts observed at that earlier state.

Here is the representation of another well-known FBT.

6.2.9. Example (Second-order FBT: the *chocolate* task). Adapted from Flobbe et al. (2008), the task is as follows. ***(i)*** Mary and John are in a room, with a chocolate bar in the room's table. ***(ii)*** John puts the chocolate into the drawer, and then ***(iii)*** he leaves the room. ***(iv)*** Mary transfers the chocolate to the box. ***(v)*** John peeks into the room, without Mary noticing, and sees the chocolate in the box. Successful performance requires that the subjects answer correctly questions such as *"where does Mary believe John will look for the chocolate?"*. This is a *second-order* task as the successful answer requires a correct *second-order* attribution.

The TVM representing this task, displayed in Figure 6.2, can be built stepwise. The initial situation is represented by s_1 (*black* indicates the chocolate (*cho*) is on the table), and each subsequent action adds a state: John (J) produces s_2 by putting the chocolate into the drawer (*white*), and he leaving the room produces s_3. Mary (M) creates s_4 when she moves the chocolate to the box (*green*), and finally s_5 emerges when John peeks into the room. In the full model:

- $M \Vdash \mathrm{B}_M(cho \lhd green) \wedge \mathrm{B}_J(cho \lhd green)$, i.e. both Mary and John believe that the chocolate is in the box, because they can both see it.

- $M \Vdash \mathrm{B}_M \, \mathrm{B}_J(cho \lhd white) \wedge \mathrm{B}_J \, \mathrm{B}_M(cho \lhd green)$, i.e. (i) Mary believes that John believes that the chocolate is in the drawer, (ii) John believes that Mary believes that the chocolate is in the box, because he *can* see her seeing the object.

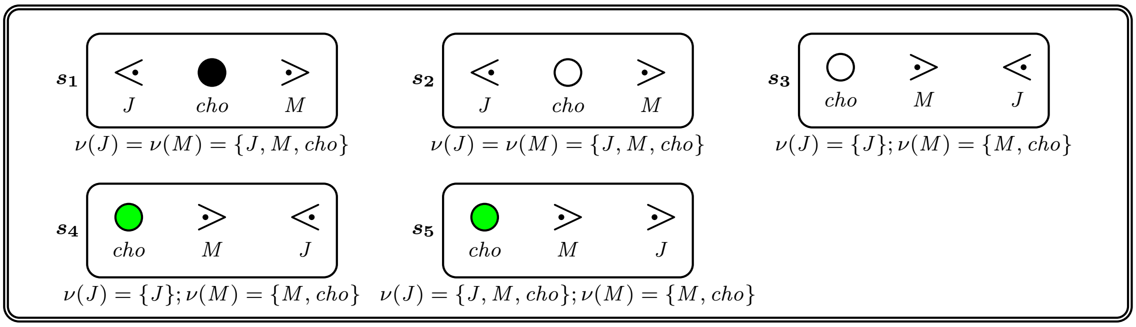

s_1
J
cho
M
$\nu(J) = \nu(M) = \{J, M, cho\}$
s_2
J
cho
M
$\nu(J) = \nu(M) = \{J, M, cho\}$
s_3
cho
M
J
$\nu(J) = \{J\}; \nu(M) = \{M, cho\}$
s_4
cho
M
J
$\nu(J) = \{J\}; \nu(M) = \{M, cho\}$
s_5
cho
M
J
$\nu(J) = \{J, M, cho\}; \nu(M) = \{M, cho\}$

Figure 6.2: TVM representation of the Chocolate Task.

- $M \Vdash \mathrm{B}_M\, \mathrm{B}_J\, \mathrm{B}_M(cho \triangleleft white) \wedge \mathrm{B}_J\, \mathrm{B}_M\, \mathrm{B}_J(cho \triangleleft white)$, i.e. (i) Mary believes that John believes that Mary believes that the chocolate is in the drawer, (ii) John believes that Mary believes that John believes that the chocolate is in the drawer, because of the last time (s_2) the visibility condition was satisfied.

6.2.2 Some technical results

As the previous subsection shows, the TV setting can represent familiar FBTs in a natural way (Section 6.2.3 discusses this in detail). We now move to a discussion of the framework's technical features, focusing on two important aspects: an alternative perspective of the proposed system (establishing connections with well-known modal logics) and a suitable notion of bisimulation (making more precise the expressivity of the chosen language). The discussion will take place in two stages, first focusing on the static fragment (leaving out model operations and their associated modalities), and then incorporating the dynamics on colour and visibility change.

The static fragment

A modal perspective Readers familiar with modal logic (Blackburn et al., 2001) will have noticed that a TVM is actually a finite linear temporal structure. Thus, it can also be described by more standard modal languages. This will be made precise now, in order to make explicit what the semantic evaluation of mental attribution formulas boils down to. For this analysis, we focus on with $\mathcal{L}'$: the fragment of $\mathcal{L}$ that does not include the model transformation modalities $[p_1{:=}c_1, \ldots, p_k{:=}c_k]$ and $[b_1{\leftarrow}X_1, \ldots, b_h{\leftarrow}X_h]$.

A modal language for describing a TVM requires special atoms for agents' visibility and objects' colour. Then, evaluating mental attribution formulas might require visiting previous states, so temporal operators are needed. A suitable one for expressing what mental attribution formulas encode is the *since* operator

$\mathrm{S}(\phi, \psi)$ (Kamp, 1968), and more precisely, its *strict* version (found also in, e.g. Burgess (1982)), read as *"since ϕ was true, ψ has been the case"*.[10]

6.2.10. DEFINITION (*Since* operator). Given a linear structure $\mathsf{M} = \langle W, \prec, V\rangle$ (with the *temporal* relation $\prec$ a strict total order[11]) and a world $w \in W$, the semantic interpretation of $\mathrm{S}(\phi, \psi)$ is as follows.

$(\mathsf{M}, w) \Vdash \mathrm{S}(\phi, \psi)$ *iff*$_{def}$ there is $u \in W$ with ***(i)*** $u \prec w$, ***(ii)*** $(\mathsf{M}, u) \Vdash \phi$, and ***(iii)*** $(\mathsf{M}, v) \Vdash \psi$ for every $v \in W$ such that $u \prec v \prec w$.[12]

Here are the formal details for this translation of the TVM framework into a temporal one. On the semantic side,

6.2.11. DEFINITION (Derived linear structure). Let $M = s_1 \dots s_n$ be a TVM with A the set of agents, O the set of objects, and R_o the set of possible values for each object $o \in O$. Define the set of atoms $P_{A,O,R_o} := \{\lhd_a x \mid a \in A, x \in A \cup O\} \cup \{o \lhd c \mid o \in O \text{ and } c \in R_o\}$. The linear structure $\mathsf{M}_M = \langle W_M, \prec_M, V_M\rangle$ over P_{A,O,R_o} consists of ***(i)*** domain $W_M := \{w_s \mid s \text{ occurs in } M\}$, ***(ii)*** temporal relation $\prec_M := \{(w_{s_i}, w_{s_j}) \in (W_M \times W_M) \mid s_i < s_j\}$ and ***(iii)*** atomic valuation $V_M(\lhd_a x) := \{w_s \in W_M \mid x \in s.\nu(a)\}$ and $V_M(o \lhd c) := \{w_s \in W_M \mid s.\kappa(o) = c\}$.

On the syntactic side,

6.2.12. DEFINITION. Formulas of the modal language $\mathcal{L}_{\mathrm{S}}$ over P_{A,O,R_o} are given by:

$$\phi ::= \lhd_a x \mid o \lhd c \mid \neg\phi \mid \phi \wedge \phi \mid \mathrm{S}(\phi, \phi)$$

for $a \in A$, $x \in A \cup O$ and $c \in R_o$. For their semantic interpretation over *pointed* linear structures, atoms $\lhd_a x$ and $o \lhd c$ are interpreted in the natural way, Boolean operators are interpreted as usual, and the *since* modality is interpreted as in Definition 6.2.10.

Finally, here is the correspondence.

[10] Note: a single "predecessor" modality is insufficient, as the number of steps the backwards exploration requires is *a priori* unknown. A modality for its reflexive and transitive closure is still not enough: it takes care of the recursive search for a state satisfying the visibility condition, but on its own cannot indicate that every state up to that point should *not* satisfy it.

[11] That is, asymmetric, transitive and total.

[12] Within propositional dynamic logic (Harel et al., 2000), and in the presence of the converse $\succ$, the *since* modality can be defined as $\mathrm{S}(\phi, \psi) := \langle(\succ; (?\phi \cup ?(\neg\phi \wedge \psi)))^{+}\rangle\, \phi$, with "?" indicating relational test, ";" indicating sequential composition, "$\cup$" indicating non-deterministic choice, and "$^{+}$" indicating one or more iterations.

6.2.13. Proposition. *Given a mental attribution formula* $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$ *in* $\mathcal{L}'$*, define the* $\mathcal{L}_{\mathrm{S}}$*-formula*

$$\mathrm{vis}_{a_1 \cdots a_m o} := \lhd_{a_1} a_2 \wedge \cdots \wedge \lhd_{a_{m-1}} a_m \wedge \lhd_{a_m} o,$$

expressing the former's visibility condition. Define the translation $tr : \mathcal{L}' \to \mathcal{L}_{\mathrm{S}}$ *as*

$$tr(\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)) := \bigvee \left\{ \begin{array}{l} \mathrm{vis}_{a_1 \cdots a_m o} \ \wedge \ o \lhd c, \\ \neg\, \mathrm{vis}_{a_1 \cdots a_m o} \ \wedge \ \mathrm{S}(\mathrm{vis}_{a_1 \cdots a_m o} \wedge o \lhd c, \neg\, \mathrm{vis}_{a_1 \cdots a_m o}) \end{array} \right\},$$

$$tr(\neg \phi) := \neg tr(\phi),$$

$$tr(\phi \wedge \psi) := tr(\phi) \wedge tr(\psi).$$

Then, for any TV *model* M *and any* $\phi \in \mathcal{L}'$,

$$M \Vdash \phi \qquad \textit{iff} \qquad (\mathsf{M}_M, w_{s_{last(M)}}) \Vdash tr(\phi).$$

Proof:
Let M be $s_1 \cdots s_n$. The proof, by induction on $\mathcal{L}'$, relies on the case for mental attribution formulas $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$. The $\mathcal{L}_{\mathrm{S}}$-formula $tr(\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c))$ holds in $\mathsf{M}_{s_1 \cdots s_n}$ at w_{s_n} iff, either the visibility condition holds and the object has the indicated colour ($\mathrm{vis}_{a_1 \cdots a_m o} \wedge o \lhd c$), or else the visibility condition fails ($\neg\, \mathrm{vis}_{a_1 \cdots a_m o}$) and there is an earlier world where both visibility and colour were satisfied, and since then visibility has failed ($\mathrm{S}(\mathrm{vis}_{a_1 \cdots a_m o} \wedge o \lhd c, \neg\, \mathrm{vis}_{a_1 \cdots a_m o})$). This is exactly what the semantic interpretation of the $\mathcal{L}'$-formula $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$ in $s_1 \cdots s_n$ requires. □

Bisimilarity The translation *tr* provides an insight on the semantic clause for mental attribution formulas. Equally illuminating is a notion of bisimilarity for $\mathcal{L}'$. One could try to find such a notion by relying on its modal *since*-based translation (see, e.g., the several notions of bisimulation discussed by Kurtonina and De Rijke (1997)). However, when describing TVMs, the language $\mathcal{L}_{\mathrm{S}}$ is clearly more expressive than $\mathcal{L}'$ (e.g. it allows nested *since* operators). Thus, although a bisimulation for $\mathcal{L}_{\mathrm{S}}$ would guarantee agreement with respect to formulas in $\mathcal{L}'$, the notion would be too strong, and agreement with respect to $\mathcal{L}'$ might not be enough to guarantee $\mathcal{L}_{\mathrm{S}}$-bisimilarity.

Still, following ideas from the literature, one can provide an appropriate notion of bisimilarity for $\mathcal{L}'$ over TVMs. Let **TVM** denote the class of all TVMs over the fixed sets A, O and $\{R_o \mid o \in O\}$.

6.2.14. Definition (TV-bisimilarity). A TV-*bisimulation* is a non-empty relation $Z \subseteq \mathbf{TVM} \times \mathbf{TVM}$ such that, for every $M = s_1 \cdots s_n$ and $M' = s'_1 \cdots s'_{n'}$ satisfying $(M, M') \in Z$, the following statement holds.

(atom) For every mental attribution formula $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$, there is $s_i \in M$ such that ***(i)*** $\mathrm{vis}_\chi(s_i)$ holds, ***(ii)*** $s_i.\kappa(o) = c$ and ***(iii)*** $\mathrm{vis}_\chi(s_k)$ fails for every $s_k \in M$ with $s_i < s_k$ iff there is $s'_j \in M'$ such that ***(i)*** $\mathrm{vis}_\chi(s'_j)$ holds, ***(ii)*** $s'_j.\kappa(o) = c$ and ***(iii)*** $\mathrm{vis}_\chi(s'_h)$ fails for every $s'_h \in M'$ with $s'_j < s'_h$.

Two TV models M and M' are said to be TV-*bisimilar* (notation: $M \underline{\leftrightarrow} M'$) iff there is a TV-bisimulation Z with $(M, M') \in Z$.

Readers familiar with Modal Logic will notice several differences between the concept provided above and the standard notion of modal bisimulation (Blackburn et al., 2001, Chapter 2). For example, while a modal bisimulation is a relation between two *pointed* relational models, a TV-bisimulation simply connects two TVMs. Moreover, a modal bisimulation consists not only of an "atom" clause, but also of two others, typically called "forth" and "back".

The reason for these differences is not only that modal formulas are evaluated in *pointed models* (pairs consisting of a model and an evaluation point), but also that the modal language contains an operator that changes the structure in which formulas are evaluated (recall: $\Diamond$ changes the pointed model by changing the evaluation point). Thus, a modal bisimulation should connect a pointed model to a pointed model while also guaranteeing that a change in one of the connected pointed models can be matched by a change in the other. In the TV framework, formulas are always evaluated with respect to the last state of the given TVM; moreover, the language $\mathcal{L}'$ has no operators for changing the structures where formulas are evaluated. Thus, a relation between TVMs is enough, and no back and forth conditions are needed. The only condition, **atom**, makes sure that the models coincide at the atomic level, i.e. with respect to mental attribution formulas. This difference only comes in support of the desired contrast between our TV models and the Kripke models used in EL. The former encode the bare minimum of information, so the TV-bisimulation only requires the **atom** clause. The latter encode all epistemic alternatives in order for the mental state operators to be interpreted as normal modal operators – hence the additional clauses required for a modal bisimulation.

It is straightforward to see that TV-bisimilarity is an alternative characterization of equivalence with respect to formulas in $\mathcal{L}'$.

6.2.15. Proposition. *For any two* TVM*s M and M', write $M \leftrightsquigarrow_{\mathcal{L}'} M'$ iff the models agree on the truth value of every $\phi \in \mathcal{L}'$. Then,*

$$M \underline{\leftrightarrow} M' \qquad \textit{iff} \qquad M \leftrightsquigarrow_{\mathcal{L}'} M'$$

Proof:
It follows from the fact that the left-hand side ($M \underline{\leftrightarrow} M'$) holds iff **atom** holds, that is, iff both models agree in the truth-value of all mental attribution formulas

$\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c)$. The right-hand side ($M \leftrightsquigarrow_{\mathcal{L}'} M'$) also requires agreement with respect to Boolean combinations of mental attribution formulas, but this follows from agreement at the atomic level. □

The crucial property of a TV-bisimulation highlights some interesting features of the TV framework. For example note how, as the **atom** clause shows, the colour of an object is relevant only if some agent can see it. Note also how two TVMs satisfying the same $\mathcal{L}'$-formulas might differ in their cardinality, and also make the same formula true in different ways (e.g. $\neg\,\mathrm{B}_a(o{\lhd}c)$ holds in M because, at $s_{last(M)}$, agent a sees o having a colour other than c, but it holds in M' because, as far as M' is concerned, agent a has never seen o).

Also interesting is to notice how, although TV-bisimilarity implies $\mathcal{L}'$-equivalence, it does not imply $\mathcal{L}$-equivalence. Take $A = \{a\}$ and $O = \{o\}$, with s_1 a state in which a sees o being white, and s_2 one in which a does not see o. Take $M = s_1$ and $M' = s_1 s_2$. The models are TV-bisimilar, hence $\mathcal{L}'$-equivalent; yet, they can be distinguished by the formula $[o{:=}black]\,\mathrm{B}_a(o{\lhd}black)$ (true in M, as the agent sees the object changing colours; false in M', as the agent does not). The different reasons why $\mathcal{L}'$-formulas are made true in TV-bisimilar models become salient when actions enter the picture.

The full framework

A dynamic modal perspective When extending the translation of Proposition 6.2.13 to formulas of the full language $\mathcal{L}$, it is enough to specify the way a (note: *finite*) derived linear structure is affected by the actions of colour and visibility change. For this, straightforward reinterpretations of the model operations for colour and visibility change are enough.

6.2.16. Definition. Let $\mathsf{M}_M = \langle W_M, \prec_M, V_M \rangle$ be a (finite) derived linear structure, with M its source TVM (Definition 6.2.11).

- Let C be a colour assignment. The structure $\mathsf{M}_M^C = \langle W_M \cup \{w_{new}\}, \prec'_M, V'_M \rangle$ expands M_M with a new world w_{new} placed at the end of the original temporal relation $\prec_M$. The atoms w_{new} satisfies reflect the colouring and visibility functions of the state s_{new} that an action of colour change with C adds to M (Definition 6.2.3).

- Let V be a visibility assignment. The structure $\mathsf{M}_M^V = \langle W_M \cup \{w_{new}\}, \prec'_M, V'_M \rangle$ expands M_M with a new world w_{new} placed at the end of the original temporal relation $\prec_M$. Again, the atoms w_{new} satisfies reflect the colouring and visibility functions of the state s_{new} that an action of visibility change with V adds to M (Definition 6.2.4).

The operations' effect can be expressed in a language $\mathcal{L}_{\mathrm{S}}^{+}$, which extends $\mathcal{L}_{\mathrm{S}}$ with additional modalities $[\alpha]$, for α a colour or a visibility assignment:

$$(\mathsf{M}_M, w) \Vdash [\alpha]\,\phi \qquad \textit{iff}_{\,def} \qquad (\mathsf{M}^{\alpha}_M, w) \Vdash \phi.$$

Thus,

6.2.17. Proposition. *Let $tr : \mathcal{L} \to \mathcal{L}^+_{\mathrm{S}}$ be a translation extending that of Proposition 6.2.13 with the clause $tr([\alpha]\,\phi) := [\alpha]\,tr(\phi)$. Then, for any* TV *model M and any $\phi \in \mathcal{L}$,*

$$M \Vdash \phi \qquad \textit{iff} \qquad (\mathsf{M}_M, w_{s_{last(M)}}) \Vdash tr(\phi).$$

Bisimilarity It has been noticed that, although TV-bisimilarity implies $\mathcal{L}'$-equivalence, it does not imply $\mathcal{L}$-equivalence. The reason is that, different from $\mathcal{L}'$, the dynamic language $\mathcal{L}$ has operators that change the structure in which formulas are evaluated. This is a problem for a TV-bisimulation, which lacks the requirements to guarantee that a change in one of the TV-bisimilar models can be matched by a change in the other.

By using ideas from Areces et al. (2015) and Aucher et al. (2018) in the context of relation-changing model operations, it is possible to obtain a notion of bisimulation that guarantees equivalence up to formulas in $\mathcal{L}$.

6.2.18. Definition ($\mathbf{R}$-based TV-bisimilarity). Let $\mathbf{R} = \{\mathsf{R}_\iota \subseteq \mathbf{TVM} \times \mathbf{TVM} \mid \iota \in I\}$ be a family of relations between TVMs. A $\mathbf{R}$-*based* TV-*bisimulation* is a non-empty relation $Z \subseteq \mathbf{TVM} \times \mathbf{TVM}$ such that, for every $M = s_1 \cdots s_n$ and $M' = s'_1 \cdots s'_{n'}$ satisfying $(M, M') \in Z$, the following statements hold.

(atom) As in Definition 6.2.14.

($\mathbf{R}$-transition) For every $\iota \in I$,

- **(ι-forth)** If there is $N \in \mathbf{TVM}$ such that $(M, N) \in \mathsf{R}_\iota$, then there is $N' \in \mathbf{TVM}$ such that $(M', N') \in \mathsf{R}_\iota$ and $(N, N') \in Z$.
- **(ι-back)** Vice versa.

Two TV models M and M' are said to be $\mathbf{R}$-TV-*bisimilar* (notation: $M \underline{\leftrightarrow}_{\mathbf{R}} M'$) iff there is a $\mathbf{R}$-based TV-bisimulation Z with $(M, M') \in Z$.

An $\mathbf{R}$-based TV-bisimulation looks not only for two TVM to coincide at the atomic level, but also for them to have matching transitions for each relation in $\mathbf{R}$. When working with the language $\mathcal{L}$, the transitions that matter are the functional ones that take a TVM to the one that results from either a colour change or else a visibility change.

6.2.19. Definition (Family $\mathbf{R}_{\mathsf{C},\mathsf{V}}$).

- Let $\mathtt{C}$ be the set of all proper colour assignments (see Definition 6.2.3) over O and $\{R_o \mid o \in O\}$. For each $C \in \mathtt{C}$, the relation $\mathsf{R}_C \subseteq (\mathbf{TVM} \times \mathbf{TVM})$ is given by $\mathsf{R}_C := \{(M, M_C) \mid M \in \mathbf{TVM}\}$.

- Let $\mathtt{V}$ be the set of all proper visibility assignments (see Definition 6.2.4) over A and O. For each $V \in \mathtt{V}$, the relation $\mathsf{R}_V \subseteq (\mathbf{TVM} \times \mathbf{TVM})$ is given by $\mathsf{R}_V := \{(M, M_V) \mid M \in \mathbf{TVM}\}$.

The family of relations $\mathbf{R}_{\mathtt{C},\mathtt{V}}$ is defined as

$$\mathbf{R}_{\mathtt{C},\mathtt{V}} := \{\mathsf{R}_C \mid C \in \mathtt{C}\} \cup \{\mathsf{R}_V \mid V \in \mathtt{V}\}$$

6.2.20. Proposition. *For any two* TVMs *M and M', write $M \leftrightsquigarrow_{\mathcal{L}} M'$ iff M and M' agree on the truth value of every $\phi \in \mathcal{L}$. Then,*

$$M \underline{\leftrightarrow}_{\mathbf{R}_{\mathtt{C},\mathtt{V}}} M' \qquad \textit{iff} \qquad M \leftrightsquigarrow_{\mathcal{L}} M'$$

Proof:
From left to right, the proof is by induction on formulas in $\mathcal{L}$. For the base case, the **atom** clause guarantees that M and M' coincide in every mental attribution formula $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$. The cases for negation and conjunction follow immediately from their respective inductive hypotheses. For dynamic formulas, let Z be the $\mathbf{R}_{\mathtt{C},\mathtt{V}}$-based TV-bisimulation satisfying $(M, M') \in Z$; suppose $M \Vdash [\alpha]\,\phi$ for α in $\mathtt{C} \cup \mathtt{V}$. Then, $M_\alpha \Vdash \phi$ but also $(M, M_\alpha) \in \mathsf{R}_\alpha$. Thus, by α-**forth**, there is $N' \in \mathbf{TVM}$ satisfying both $(M', N') \in \mathsf{R}_\alpha$ and $(M_\alpha, N') \in Z$. From the latter, $M_\alpha \Vdash \phi$ and inductive hypothesis, $N' \Vdash \phi$; then, the determinism of R_α implies $N' = M'_\alpha$. Therefore, $M'_\alpha \Vdash \phi$, that is, $M' \Vdash [\alpha]\,\phi$. The direction from $M' \Vdash [\alpha]\,\phi$ to $M \Vdash [\alpha]\,\phi$ works analogously, using α-**back** instead.

From right to left, it is enough to show that $\leftrightsquigarrow_{\mathcal{L}}$ is a $\mathbf{R}_{\mathtt{C},\mathtt{V}}$-based TV-bisimulation, so suppose $M \leftrightsquigarrow_{\mathcal{L}} M'$. Clearly, **atom** holds, as M and M' satisfy the same mental attribution formulas. For α-**forth**, take any $\alpha \in \mathtt{C} \cup \mathtt{V}$ and the unique M_α. For a contradiction, suppose there is no $N' \in \mathbf{TVM}$ satisfying both $(M', N') \in \mathsf{R}_\alpha$ and $M_\alpha \leftrightsquigarrow_{\mathcal{L}} N'$. Since R_α is functional, there is exactly one $M'_\alpha \in \mathbf{TVM}$ satisfying $(M', M'_\alpha) \in \mathsf{R}_\alpha$, so the problem should be that $M_\alpha \not\leftrightsquigarrow_{\mathcal{L}} M'_\alpha$. But then there would be a formula ϕ in whose truth-value M_α and M'_α differ, and thus M and M' would differ in the truth-value of $[\alpha]\,\phi$, contradicting $M \leftrightsquigarrow_{\mathcal{L}} M'$. Hence, α-**forth** holds. The argument for α-**back** is analogous. □

Still, these $\mathbf{R}$-based TV-bisimulation are a bit abstract, and for some purposes one might be interested in more concrete insights. For example, when considering colour change, is there an alternative characterization of a $\{\mathsf{R}_C \mid C \in \mathtt{C}\}$-based TV-bisimulation? Since colour change essentially affects the attributions for agents who can see the objects in question, a candidate relation is the relation $Z \subseteq (\mathbf{TVM} \times \mathbf{TVM})$ satisfying **atom** and the following clause

- for all $a \in A$ and $o \in O$,

$$o \in s_{last(M)}.\nu(a) \quad \textit{iff} \quad o \in s'_{last(M')}.\nu(a).$$

A study of these ideas is left for future work.

6.2.3 Evaluating the framework

The TV framework allows us to model FBTs where belief attributions are formed on basis of visibility and recollection of earlier events. So far, it has been used for representing two paradigmatic FBTs.

Looking back at the desiderata of Section 6.1, the approach fulfils *robustness*: it can be uniformly applied to different scenarios, regardless of the order of the involved mental attributions, witness the *Sally-Anne Task* (first-order) and the *Chocolate Task* (second-order). Moreover, *faithfulness* is also fulfilled because events as those involved in the described scenarios (moving an object, coming to (un)see entities) correspond naturally to the colour and visibility changes accounted for in our syntax and semantics.[13]

There is another criterion we set for this modelling attempt: to gain insights on why people might fail FBTs. Consider, for example, the Sally-Anne Task. At the end of the story, Sally believes the marble is in the basket (i.e. $M \Vdash \mathrm{B}_{Sa}(mar \triangleleft white)$), therefore having a false belief.[14] Anne "knows" this, as she believes that Sally believes the marble is in the basket: $\mathrm{B}_{An}\, \mathrm{B}_{Sa}(mar \triangleleft white)$ holds at M. The modelling thus produces the correct answers, as a modelling within standard EL would do.[15] Yet, the semantic interpretations of our formulas rely on visibility and memory, two ingredients that play a large role on the limitations of human reasoning, as indicated by empirical research. By making explicit the role they play, the system allows us to represent why Anne's reasoning process *might* fail.

For example, one explanation for potential failures has to do with the cognitive difficulty of making very involved belief attributions. Indeed, research suggests that human working memory has a capacity of around 4 "chunks" (Cowan, 2001). In the TV framework, this human limitation can be brought to the table by simply

[13] Notice that other actions taking place in other FBTs, e.g. involving communication, will be addressed in the next section.

[14] Having a false belief does not imply that there is "something wrong" with Sally. In game-theoretical terms (e.g. (Osborne and Rubinstein, 1994, Chapter 11)), she is simply an agent with *imperfect information*, that is, an agent that has not been informed of all the relevant actions that have taken place. In real-life, false beliefs are common, and are often used by epistemologists to discern belief from knowledge, which is supposed to be factive. In the TV framework, false beliefs arise because agents do not need to have full direct visibility of the involved agents/objects in the most recent state. In these cases, they have to recall earlier states in which the needed information was available, even though the situation might have changed since.

[15] Our framework can also evaluate attributions of any length, as in possible-worlds semantics.

limiting the number of states an agent is able to go "back in time", or even the number of agents whose visibility she can keep track of. Another source of failures in attributing beliefs has to do with the *type of agents* making the attribution. For example, an autistic agent might fail to attribute beliefs to others correctly, instead reporting her own (Baron-Cohen et al., 1985). The framework can be fine-tuned to capture agents with a special type of higher-order reasoning, e.g. through the introduction of a *perceived agent* function $\pi : A \to (A \to A)$, with $\pi_a(b) = c$ understood as *"agent a considers agent b to have the perspective of agent c"*). This can be then used to define an appropriate variation of the visibility condition, with a not looking for what b can see, but rather looking for what *the agent she perceives as b* (i.e. $\pi_a(b)$) can see. In this way, an autistic agent a would be one for which $\pi_a(x) = a$ for any $x \in A$, essentially relying only on her own information, and thus attributing her own belief to others.

Perhaps more interestingly, the setting can also bring to light other reasons why belief attributions might fail. The interpretation of mental attribution formulas relies not only on the size of the agent's memory but also on its *accuracy*. Thus, if the recollection of an agent has gaps or lacks the correct order, the backtracking will naturally fail. In other words, even agents with a large memory capacity might make incorrect attributions, as their recollection of earlier states might be chronologically incorrect. All these possible sources of failure can be tested experimentally, in order to determine whether they play a role in the difficult task of making higher-order attributions. In turn, the findings can inform the design of resource-bounded formal frameworks, whereby agents simply "give up" in making those attributions that exceed the empirically indicated bound on a cognitive resource (like working memory). This is in agreement with the view spelled out in Section 2.5.2 regarding the contribution of logical modelling to the generation of testable hypotheses and experimental designs.

Further cognitive features

The TV framework also reveals further features thought of as crucial ingredients of social cognition, often contrasted with EL modelling.

Informational economy On the one hand, a state in a TVM contains a bare informational "minimum": only basic facts regarding objects and agents' visibility. The operations on the model also induce minimal changes, in accordance with the criterion of informational economy in belief revision (Gärdenfors, 1988). On the other hand, the non-standard semantic clause for belief is complex, as the state representing the current situation might not have all information necessary to evaluate a complex belief attribution, and thus the information at other (previous) stages might be needed.[16] A "backtracking" process might be difficult and time-consuming, depending on how many different states an agent needs to

[16] Tracking the reasons why an agent forms a belief in the semantic interpretation might be helpful from the epistemologist's point of view as well. Revising EL to capture more fine-grained

remember, and our clause is sensitive to this observation, unlike the usual modal interpretations. The level of complexity that one finds on this framework for both representing a situation (low) and evaluating mental attributions (high) can be contrasted with what EL does, as discussed in Section 6.1.

Perspective shifting Another important feature, identified in analyses of ToM and formalizations of FBTs, is *perspective shifting* (Braüner, 2014). Successful performance in the tasks (i.e. making correct attributions) requires a perspective shift: stepping into the shoes of another agent.[17] The evaluation of mental attribution formulas does not ask for an explicit shift in perspective (as, e.g. the repetitive calling of a recursive function requires). Yet, the evaluation forces the main agent in the attribution (i.e. agent a_1 in each attribution $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$) to consider the visibility of the other agents (or, in terminology of the next section, to consider what they *perceive* is the visibility of the other agents). The difference between the different perspectives is what might force the main agent to recall earlier stages in order to evaluate the attribution. The more agents are involved (i.e. the more complicated the attribution is), the more complex the visibility condition becomes, capturing in this way why agents may sometimes fail the tasks.

Principle of inertia A further crucial notion is the *principle of inertia* (Stenning and van Lambalgen, 2008; Braüner, 2015; Braüner et al., 2016): an agent's beliefs are preserved unless there is reason to the contrary. In our case, reason to the contrary amounts to the satisfaction of visibility; if this is not satisfied in the state of evaluation, then, essentially, the agent maintains beliefs formed in earlier stages, where necessary information was available. Inertia can be seen as an "economical" policy in forming (higher-order) beliefs, especially in the context of the tasks because the experimenters design structured and controlled storylines, facilitating straightforward responses.[18]

Connections with the dual process theories of reasoning Besides ToM and FBTs, the presented setting can also draw connections between logical formalization and dual process theories of reasoning (Section 2.3.4). We argue that agents' higher-order reasoning roughly follows this pattern. System 1 keeps track only of a bare-minimum of information (basic facts), without overloading memory with

notions of belief, e.g. notions that include its justifications as well, is a common strategy in attempts that use logic to inform epistemological debates on theories of knowledge.

[17] In fact, unsuccessful performance, e.g. of autistic children, is often connected with a failure in perspective shifting, resulting in the subject reporting her own beliefs (Stenning and van Lambalgen, 2008; Braüner, 2015).

[18] Of course, people's general use of ToM would not always adhere to such a principle: different agents might be naturally more suspicious about what happens during their absence, and lack of visibility of objects/agents in question might last long enough for them to become hesitant and refrain from using inertia. But even in those cases, the TV framework might be of use, changing the idea of the *visibility* condition for the more general idea of an *evaluation* condition, indicating what a state needs to satisfy in order to be used for evaluating a mental attribution.

information that can be later inferred. Whenever a task requires more than what is stored (as higher-order attributions), System 2 takes over, using the inputs of System 1. This is precisely the pattern of our semantics, with our models and updates encoding only basic facts. Whenever a demanding task appears, such as the evaluation of a mental attribution, our agents follow the cognitively hard calculations of our semantic clause. On the basis of elementary facts regarding whom/what they observed, they test certain conditions and trace back earlier states. It is only after this slow and effortful process that they can determine whether a higher-order attribution holds.

6.2.4 Related proposals

The TV framework can be compared with other proposals. For the purposes of this text, the most meaningful comparisons are with other approaches for studying ToM and representing FBTs. However, it is also worthwhile to mention the connections with logical frameworks making use of visibility and memory, the two fundamental aspects the TV framework relies on.

Representing mental attributions Let's start with an overview of attempts in the first direction.

In one of the first EL-based proposals for dealing with ToM, van Ditmarsch and Labuschagne (2007) study three kinds of agents (including agents on the autism spectrum) endowed with specific "strategies" they use for higher-order reasoning. They use relational preference structures for modelling different degrees of belief, similar to the ones used by Board (2004); van Benthem (2007); Baltag and Smets (2008b) when dealing with (relational) belief revision. In contrast, our attempt does not focus on agents with specific strategies when evaluating belief attributions; it rather considers *any* agent's reasoning behind such processes. This is closely related to our choice of robustness as an evaluative criterion for the framework. Still, as discussed in Section 6.2.3, here agents can be arranged by their cognitive capacities, as the size of their memory or the way they perceive others. This, together with their visibility, roughly decides the way they will make belief attributions.

Using different tools, Stenning and van Lambalgen (2008) provide instead a non-monotonic closed-world reasoning formalization of first-order FBTs, implemented within logic programming. They use *event calculus* (van Lambalgen and Hamm, 2008), with belief treated as a predicate, and rely on the principle of inertia. While we design a different formalism, we still account for these features without restricting ourselves to specific types of agents or orders of beliefs. Again, this is to ensure that the framework is robust enough to model a wide variety of FBTs in a uniform way.

Another interesting logical formalization of FBTs is given by Braüner (2015) and Braüner et al. (2016). These papers use a proof-theoretic Hybrid Logic system

(Braüner, 2017) for the analysis of well-known FBTs, identifying perspective shifts and using inertia. The most straightforward difference is that our approach is rather semantic, with models keeping track of the actions involved, and in which the evaluation of mental attributions reflects their cognitive difficulty. Some of the tasks formalized in the proof-theoretic analysis involve communication, in particular misinformation; we will present an extension of our semantic model to accommodate misinformation in the next section. An interesting feature of this analysis is that it does not only formalize what happens when subjects give the correct responses to the task questions: it can also identify what happens when incorrect responses are given. Our approach has focused on the correct attributions, yet we have discussed how it can indicate possible sources of failure.

Bolander (2018) uses (D)EL tools (models and language) plus special atoms indicating the location of objects and the agents' visibility. Then, it represents changes in the situation as action-model-based acts of (private) communication (Baltag et al., 1998) that rely on agents' visibility.[19] The most important differences between our proposal and this (and other frameworks relying on EL) have been already discussed: the contrast between complex models that simplify answering mental attribution questions (EL) and simple states that require a complex process for deciding higher-order belief issues (here). The representation of actions also differs: while Bolander (2018) uses (a variation of) the heavy action models machinery (for private communication), the actions of visibility and colour change presented here simply modify atomic information (while keeping track of the past). Finally, the TVM framework fulfils the requirements Bolander proposes: it is robust enough to deal with different FBTs, and the actions in the stories have a straightforward representation, as we have argued before.

On visibility/observability and memory Other logical systems have already emphasized the role that visibility/observability and memory play on an agent's epistemic state.

With respect to the first, the list of proposals include Charrier et al. (2016); Herzig et al. (2018); van Benthem et al. (2018), whose underlying idea that epistemic attitudes (knowledge, beliefs) are built from observation and communication is shared by the TV framework.[20] They too propose more "compact" struc-

[19] For example, the act through which, in the absence of Sally, Anne moves the marble from the basket to the box, is understood as a private announcement through which only Anne is informed about the marble's new location. Based on this reasoning system, Dissing and Bolander (2020) provide an implementation on a humanoid robot that successfully passes a general class of FBTs. Van De Pol et al. (2018) provide a *dynamic belief update* model, similar to Bolander's, to analyze the computational complexity of ToM reasoning, which they show to be intractable, but not due to the order of reasoning (which can nonetheless affect *cognitive* difficulty for reasons other than computational complexity, e.g. due to limits on working memory).

[20] In a *coalition logic* context, van der Hoek et al. (2011) also rely on the idea of agents being able to observe the truth-value of only certain atoms, combining it with the idea of agents having *control* over the truth-value of certain atomic propositions.

tures, compared to Kripke models, representing basic facts in terms of which knowledge is interpreted. However, there is a difference. In these frameworks, the Kripke models, hence the standard interpretations of mental states, can be recovered from the alternative structures and interpretations. Our alternative structures, the TV models, are designed to avoid this equivalence, given our motivation for studying realistic mental state attribution. We seek an alternative modelling to reflect the reasoning underlying attributions real people make in experimental tasks, while the aforementioned approaches seek an alternative modelling to derive technical benefits, e.g. concerning symbolic model checking. This difference in motivation is precisely what explains the differences between our attempt and the aforementioned ones.

Other logical frameworks have also paid attention to the agents' memory, albeit in different ways. On the one hand, most works looking at the issue from a resource-bounded perspective have focused on the agents' *working memory*, that is, the space they have for performing a task (Albore et al., 2006; Alechina et al., 2008, 2009b). On the other hand, Liu (2009) looks at (DEL-based) agents who might forget things they knew in the past. Our understanding of memory is closer to the latter, as the model encodes not only the current state of affairs, but also the way things were before. Still, a crucial difference remains. In the listed proposals, the knowledge/beliefs of an agent depend only on information present at the current stage. In the TV framework, the beliefs of an agent depend not only on what she can see now, but also on (what she remembers about) what she could see in the past.

6.3 Multiple perspectives

The previous section showed how the TV framework can be used for modelling some FBTs (Section 6.2.1). Still, for some other tasks, a model based on visibility and memory is not enough.

6.3.1. Example (The *Bake-Sale Task*). The following is adapted from Hollebrandse et al. (2014).

> *While at home, Mary and John find out that the church is having a bake sale where chocolate cookies are sold. Mary then heads for the church to get chocolate cookies. But, after Mary leaves, their mom comes back home and tells John that, according to her, only pumpkin pie is being sold. Mary arrives at the bake sale but, in fact, all there is for sale is brownies.*

In the original story, Mary meets the mailman in her way back. The mailman asks *"Does John know what you bought him?"*. Therefore, the crucial second-order formula to be evaluated concerns what Mary believes about what John believes over the product she purchased for him.

In the Bake-Sale Task, both John and Mary receive information not only by means of observation; they also get it through some other agents. This has an important consequence: unlike information obtained through visibility, information obtained from other agents might not be truthful. More important: this potentially false information might affect the belief attributions agents make, much like the different ways an agent might perceive others might produce false belief attributions (van Ditmarsch and Labuschagne, 2007). For example, if an agent is told that an object is black, then she thinks she has access to the object (let's call this *perceived visibility*) and that its colour is black (let's call this *perceived colour*). While perceived and factual descriptions (about visibility and colours) can coincide, they can also diverge when misinformation and lying is involved.

It should be clear then that the TV framework of the previous section is not sufficient for representing tasks as the Bake-Sale scenario. In order to do that, the setting should be extended for dealing with the (not necessarily truthful) information the agents acquire via *communication*.

This section will provide the tools for dealing with scenarios like the Bake-Sale task. First, the TVMs of Definition 6.2.1 will be supplemented with functions for representing the agents' *subjective's perception* on the colour of objects and the visibility of agents. Then, the operations for colour and visibility change (Definition 6.2.3 and Definition 6.2.4) will be redefined for working on the new structures. Finally, additional operations will be defined for representing acts of communication. As before, let $A \neq \varnothing$ be the set of agents, $O \neq \varnothing$ the set of objects and R_o the range of possible colours of each object $o \in O$.

6.3.2. Definition (Multi-perspective TVM). A *multi-agent state* is a tuple $s = \langle \kappa, \nu, \{\kappa^a, \nu^a\}_{a \in A} \rangle$ where κ and ν are the *factual* colouring and visibility functions, and each κ^a and ν^a are the *perceived* colouring and visibility functions for agent $a \in A$. All colouring and visibility functions, both factual and perceived, should satisfy the requirements stated in Definition 6.2.1.[21] Additionally, every *perceived* visibility function ν^a is required to satisfy the following *perceived self-visibility* constraint: the agent's perceived visibility contains at least her real visibility: $\nu(a) \subseteq \nu^a(a)$.

A *multi-perspective* TVM (MPTVM) $\mathbf{M}$ is a finite non-empty sequence of multi-agent states $\mathbf{M} = s_1 \cdots s_n$. As before, $\sharp_{\mathbf{M}} := n$ is the model's cardinality, and $s_{last(\mathbf{M})} := s_{\sharp_{\mathbf{M}}}$ is its *last* state. Again, write $s_i < s_j$ iff $i < j$.

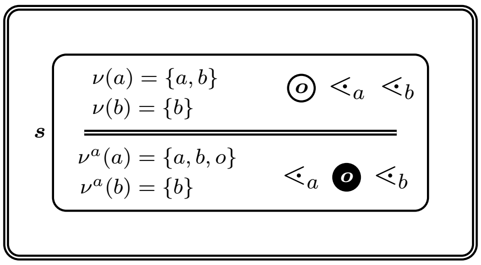
A multi-perspective TVM is then a TVM in which each agent has her own perspective about the visibility and the colouring that has occurred at each stage (i.e. state). The additional perceived self-visibility constraint guarantees that every agent thinks she sees everything she can actually see, and yet leaves the room open for her to assume she can see more things. This is because of the

[21] Thus, every object has a proper colour and every agent can see herself.

effect of acts of communication, which may make an agent think she can "see" (understood as "have access to") objects/agents only because she has received information about them.

There are other restrictions one might want to impose on a multi-agent state s, particularly about the relationship between the factual and perceived functions. For example, one might require *correct perceived colour for observed objects*, so agents perceive correctly the colour of the objects they can actually see (for every $o \in s.\nu(a)$, we have $s.\kappa^a(o) = s.\kappa(o)$). Similarly, one might require *correct perceived visibility for observed agents*, so agents perceive correctly the visibility of the agents they can actually see (for every $b \in s.\nu(a)$, we have $s.\nu^a(b) = s.\nu(b)$). Still, these requirements are not essential for what will be discussed later;[22] thus, they will not be mandatory. Here is a simple example:

6.3.3. Example. Consider an object o and two agents a and b. Let s be the following multi-agent state: the first row shows the real situation and the second one shows a's perspective (the point of view of b is not displayed here).

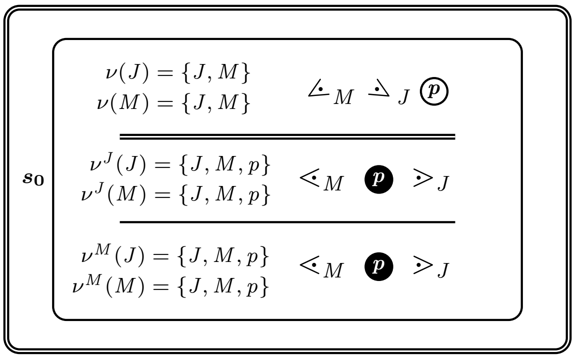


Note how $\nu^a(a) = \{a, b, o\}$, so agent a considers that she can see herself, agent b and object o. Yet, this is not the case, as in reality she can only see herself and b (as $\nu(a) = \{a, b\}$). Observe also how a thinks, incorrectly, that o has colour *black*. Here this might be explained because in fact she cannot see o; still, recall that correct perceived colour for observed objects is not being enforced, so she might be mistaken even if she can actually see it.

Going back to the Bake-Sale Task, we give another example of a MPTVM:

6.3.4. Example. In the initial stage of the Bake-Sale Task, neither John (J) nor Mary (M) can see the product sold (p), but they do see each other. They both believe that the product is chocolate cookies (*black*), because they have heard so, while in fact the product is brownies (*white*). A MPTVM capturing this is given below, with its single state s_0 one with not only $\kappa(p) = \textit{white}$ and $\kappa^J(p) = \kappa^M(p) = \textit{black}$, but also $\nu(J) = \{J, M\} = \nu(M)$ and $\nu^J(J) = \nu^J(M) = \nu^M(J) = \nu^M(M) = \{J, M, p\}$.

[22] For example, consider correct perceived colour for observed objects, asking $\kappa^a(o) = \kappa(o)$ for every $o \in \nu(a)$. An agent's perception of an object's colour is only relevant when she "thinks" she sees the object ($o \in \nu^a(a)$); in those cases, what matters is the colour she thinks the object has ($\kappa^a(o)$), regardless of the object's actual colour ($\kappa(o)$).



Factual change in the new setting The next step is to define the way acts of factual change will affect a MPTVM. First, for colour change:

6.3.5. Definition (Colour change). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Take a set of objects $\{o_1, \ldots, o_k\} \subseteq O$, with $c_i \in R_{o_i}$ $(1 \leqslant i \leqslant k)$ a proper colour for each one.

The colour assignment $C = [o_1{:=}c_1, \ldots, o_k{:=}c_k]$ produces the MPTVM $\mathbf{M}_C = s_1 \cdots s_n s_{new}$. In the new multi-agent state s_{new}, the factual functions are given exactly as in Definition 6.2.3.[23] The action does not affect the agents' perceived visibility, which is then given at s_{new} for every $a \in A$ and every $\ell \in A$ as $s_{new}.\nu^a(\ell) := s_{last(\mathbf{M})}.\nu^a(\ell)$. Then, the perceived colouring of each $a \in A$ at s_{new} is given, for every $o \in O$, by

$$s_{new}.\kappa^a(o) := \begin{cases} c_i & \text{if } o = o_i \in \{o_1, \ldots, o_k\} \text{ and } o \in s_{last}.\nu(a) \\ s_{last(\mathbf{M})}.\kappa^a(o) & \text{otherwise} \end{cases}$$

so a "notices" the change of colour of the objects in C she could see (equivalently, *"can see"* $[o \in s_{new}.\nu(a)]$, as factual visibility in s_{new} is exactly as in s_{last}), keeping her previous perception of colour otherwise.

The resulting structure is indeed a MPTVM, as the operation preserves the three required properties.

6.3.6. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM, *with the colour assignment* C *and the structure* $\mathbf{M}_C$ *as in Definition 6.3.5. Then,* $\mathbf{M}_C$ *is a* MPTVM.

Proof:
States in $\mathbf{M}$ are already proper multi-agent states; only the new multi-agent state s_{new} is left to check. For the first requirement, s_{new} assigns *proper colours* to objects, both factually and perceivably: for objects affected because of the condition on each c_i, and for the rest because they inherit their (proper) colours

[23] Thus, while factual visibility in s_{new} is exactly as in $s_{last(\mathbf{M})}$, factual colouring in s_{new} takes the colouring of $s_{last}(\mathbf{M})$ for objects not occurring in C, following C for the objects it mentions.

from $s_{last(\mathbf{M})}$. For the second and the third requirements, factual and perceived visibility in s_{new} are inherited from $s_{last(\mathbf{M})}$. Thus, since the latter satisfied *every agent can (factually and perceivably) see herself* and *perceived self-visibility*, so does the former. □

The operation also preserves the two other discussed requirements.

6.3.7. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*, with the colour assignment* C *and the structure* $\mathbf{M}_C$ *as in Definition 6.3.5. If* $\mathbf{M}$ *satisfies* ***(i)*** correct perceived colour for observed objects*,* ***(ii)*** correct perceived visibility for observed agents*, then so does* $\mathbf{M}_C$*, respectively.*

Proof:
Again, it is enough to check that the new multi-agent state s_{new} satisfies the requirements, so let $a \in A$ be an agent. For ***(i)***, take $o \in s_{new}.\nu(a)$ (so $o \in s_{last}.\nu(a)$). If o is some $o_i \in \{o_1, \ldots, o_k\}$, both $s_{new}.\kappa(o)$ and $s_{new}.\kappa^a(o)$ are given by c_i (the second, because $o \in s_{last}.\nu(a)$); otherwise, both $s_{new}.\kappa(o)$ and $s_{new}.\kappa^a(o)$ remain as in $s_{last(\mathbf{M})}$, where they coincided because this last state satisfies the property. For ***(ii)***, it is enough to notice that both factual and perceived visibility in s_{new} are, for every $\ell \in A$, exactly as in $s_{last(\mathbf{M})}$, where they coincide. □

The just defined operation of colour change affects not only the colour of the listed objects, but also these objects' perceived colour for those agents that can see them. The operation for visibility change in MPTVMs, to be defined below, works somehow analogously: it changes factual visibility, affecting also the perspective of both the agents that now can see the affected ones, but also that of the agents that could see them before. This emphasizes the idea of all involved agents changing their visibility *simultaneously*. Additionally, the operation allows the affected agents to realize the actual colour of the objects they can now see.

6.3.8. Definition (Visibility change). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Take a set of agents $\{b_1, \ldots, b_h\} \subseteq A$, with sets $X_i \subseteq A \cup O$ $(1 \leqslant i \leqslant h)$ satisfying $b_i \in X_i$.

The visibility assignment $V = [b_1 \leftarrow X_1, \ldots, b_h \leftarrow X_h]$ produces the MPTVM $\mathbf{M}_V = s_1 \cdots s_n s_{new}$. In the new multi-agent state s_{new}, the factual functions are defined exactly as in Definition 6.2.4.[24] The perceived visibility of each agent $a \in A$ is defined, for every $\ell \in A$, as

$$s_{new}.\nu^a(\ell) := \begin{cases} X_i & \text{if } \ell = b_i \in \{b_1, \ldots, b_h\} \text{ and } \ell \in (s_{new}.\nu(a) \cup s_{last}.\nu(a)) \\ s_{last(\mathbf{M})}.\nu^a(\ell) & \text{otherwise} \end{cases}$$

[24] Thus, while factual colouring in s_{new} is exactly as in $s_{last(\mathbf{M})}$, factual visibility in s_{new} takes the visibility of $s_{last}(\mathbf{M})$ for agents not occurring in V, following V for the agents it mentions.

so a "notices" the visibility change of the agents she can or could actually see, keeping her previous perception otherwise. Finally, the perceived colouring of each agent $b \in \{b_1, \ldots, b_h\}$ is defined, for every $o \in O$, as

$$s_{new}.\kappa^b(o) := \begin{cases} s_{new}.\kappa(o) & \text{if } o \in s_{new}.\nu(b) \\ s_{last(\mathbf{M})}.\kappa^b(o) & \text{otherwise} \end{cases}$$

so agents whose visibility changes "notice" the factual colour of the objects they can see now, keeping their previous perceived colour for the rest. For the rest of the agents, perceived colouring remains exactly as before ($s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for $a \in A \setminus \{b_1, \ldots, b_h\}$).

Once again, the resulting structure is indeed a MPTVM.

6.3.9. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*, with the visibility assignment* V *and the structure* $\mathbf{M}_V$ *as in Definition 6.3.8. Then,* $\mathbf{M}_V$ *is a* MPTVM*.*

Proof:
Factual colouring in the new state s_{new} is inherited from the (proper) factual colouring in $s_{last(\mathbf{M})}$; moreover, perceived colouring is either inherited or else taken from the factual one. Thus, the new model assigns proper colours, both factually and perceivably.

For self-visibility, the condition on each X_i and the fact that $s_{last(\mathbf{M})}$ is a proper multi-agent state guarantees that, according to both factual and perceived visibility in s_{new}, every agent can see herself. For *perceived self-visibility*, take any agent $a \in A$. If her factual visibility is affected by the operation, both factual and perceived visibility become the corresponding X_i (for the second, because $a \in (s_{new}.\nu(a) \cup s_{last}.\nu(a))$), thus satisfying the property. Otherwise, both factual and perceived visibility are taken from $s_{last(\mathbf{M})}$, and so the property is inherited. □

The operation also preserves the two other discussed requirements.

6.3.10. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*, with the visibility assignment* V *and the structure* $\mathbf{M}_V$ *as in Definition 6.3.8. If* $\mathbf{M}$ *satisfies* ***(i)*** correct perceived colour for observed objects*,* ***(ii)*** correct perceived visibility for observed agents*, then so does* $\mathbf{M}_V$*, respectively.*

Proof:
Let $a \in A$ be an agent. Item ***(i)*** for agents whose visibility changes is immediate, as the definition directly establishes that, at the new state s_{new}, the colouring a perceives for objects she can see is the factual one. For the rest of the agents, the property is inherited from $\mathbf{M}$. Item ***(ii)*** also follows; take any ℓ that a can see at s_{new}: if ℓ's visibility is affected by the operation, both factual and perceived visibility become the same new set; otherwise, both factual and perceived visibility

are inherited from $s_{last(\mathbf{M})}$, where they coincide. □

Acts of communication The two just defined operations essentially rewrite the effect of factual changes in the colour of objects and the visibility of agents (additionally making sure that changes in visibility give the agents the proper perception). Now, here is a novel action the new structure allows: one through which a set of agents get informed about the colour of an object.

6.3.11. Definition (Communicating an object's colour). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Let $B \subseteq A$ be a set of agents and $p \in O$ an object, with $c \in R_p$ a proper colour.

The message $p{:=}c$ for agents in B produces the MPTVM $\mathbf{M}_{B(p:=c)} = s_1 \cdots s_n s_{new}$. In the new multi-agent state s_{new}, factual functions are inherited from the last state in $\mathbf{M}$,[25] just like the perceived functions for agents *not in* B.[26]

Then, for agents $a \in B$: their perceived colouring at s_{new} is given, for every $o \in O$, as

$$s_{new}.\kappa^a(o) := \begin{cases} c & \text{if } o = p \text{ and } p \notin s_{last}.\nu(a) \\ s_{last(\mathbf{M})}.\kappa^a(o) & \text{otherwise} \end{cases}$$

so a's perceived colour about o becomes c only if this is the involved object and she *could not* see it,[27] remaining as before otherwise. Then, their perceived visibility at s_{new} is given, for every $\ell \in A$, as

$$s_{new}.\nu^a(\ell) := \begin{cases} s_{last(\mathbf{M})}.\nu^a(\ell) \cup \{p\} & \text{if } \ell = a \\ s_{last(\mathbf{M})}.\nu^a(\ell) & \text{otherwise} \end{cases}$$

so a's perceived visibility about herself is extended with p, and her perceived visibility about all other agents remains as before.

This act of communicating colouring, just like the act of communicating visibility below, concerns information about a single entity (a single object in the first, a single agent in the second). In this sense, the communication actions are simpler than the actions of colour and visibility change of before, through which multiple entities (objects/agents) are modified. This is for simplifying the presentation of the communication operations, but with the proper adjustments they can be extended for dealing with multiple entities.

[25] More precisely, $s_{new}.\kappa(o) := s_{last(\mathbf{M})}.\kappa(o)$ for all $o \in O$, and $s_{new}.\nu(\ell) := s_{last(\mathbf{M})}.\nu(\ell)$ for every $\ell \in A$.

[26] Thus, for each $a \in A \setminus B$, we have $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for all $o \in O$, and $s_{new}.\nu^a(\ell) := s_{last(\mathbf{M})}.\nu^a(\ell)$ for every $\ell \in A$.

[27] Equivalently, *"cannot see"* ($o \in s_{new}.\nu(a)$), as factual visibility in s_{new} is exactly as in s_{last}.

Note how an act of colour communication for agents in B affects only the way *these* agents perceive the situation; for the rest, no perceptual change occurs. Moreover: those agents in B that could actually see the discussed object in the previous state[28] will not change their perception about it; this is to ensure that agents give priority to their factual visibility over communicated information. Finally, receiving information about the colour of an object makes an agent believe she can actually see the object.[29] Nevertheless, she does not change her perception about other agents' visibility, not even about the agents that also receive the message (i.e. the other agents in B). In other words, all agents in B receive the information consciously, as each one of them does two things: update her perspective about p's colour, and assume she can see p. However, no agent in B makes any assumption about whether someone else received the information: no one changes her perceived visibility about the others. Indeed, no agent in B *adds* p to what she thinks other agents can see (what she would need to do for the agents she thinks *have also received* the information), and neither removes p from what she thinks other agents can see (what she would need to do for the agents she thinks *have not received* the information).

6.3.12. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$*,* $p \in O$*,* $c \in R_p$ *and the structure* $\mathbf{M}_{B(p:=c)}$ *as in Definition 6.3.11. Then,* $\mathbf{M}_{B(p:=c)}$ *is a* MPTVM.

Proof:
For proper colouring, factual colouring in the new state s_{new} is inherited, so it assigns proper colours. Likewise, perceived colouring in s_{new} assigns proper colours when it is inherited, and also when it is new (due to $c \in R_p$). For self-visibility, every agent can see herself in s_{new} under both factual and perceived visibility: for the first because it is inherited, and for the second because it is either inherited or else extended. Finally, *perceived self-visibility* also holds at s_{new} because, while factual visibility is inherited, perceived visibility is either inherited or else extended, but never reduced. □

The operation also preserves one of the optional requirements.

6.3.13. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$*,* $p \in O$*,* $c \in R_p$ *and the structure* $\mathbf{M}_{B(p:=c)}$ *as in Definition 6.3.11. If* $\mathbf{M}$ *satisfies* correct perceived colour for observed objects*, so does* $\mathbf{M}_{B(p:=c)}$.

Proof:
Let s_{new} be the new state and $a \in A$ be an agent; take any $o \in s_{new}.\nu(a)$. Factual visibility does not change, so $o \in s_{last}.\nu(a)$. Then, from the definition, the agent's

[28] Equivalently, those agents in B that *can* actually see the discussed object in the new state. This is because *factual* visibility in both s_{new} and s_{last} is the same.

[29] Note: the action only *adds* the discussed object, keeping all her previous perceived visibility.

o's perceived colour at s_{new} is inherited from s_{last}. But factual colouring does not change either, so o's factual colour at s_{new} is inherited from s_{last}. Thus, if both o's factual colouring and a's perception of it coincided in s_{last}, they also coincide in s_{new}. □

Still, the communication operation does not preserve *correct perceived visibility for observed agents*. The reason is that communication makes the agent believe she can see the involved object, when this might actually not be the case. Since an agent can always see herself, she will not have a correct perception about her own visibility.

The action of communicating the visibility of some agent works analogously.

6.3.14. Definition (Communicating an agent's visibility). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Let $B \subseteq A$ be a set of agents and $d \in A$ an agent, with $X \subseteq A \cup O$ a set of agents and objects satisfying $d \in X$.

The message $d \leftarrow X$ for agents in B produces the MPTVM $\mathbf{M}_{B(d\leftarrow X)=s_1\cdots s_n s_{new}}$. In the new multi-agent state s_{new}, the factual functions are inherited from the last state in $\mathbf{M}$,[30] just like the perceived functions for agents *not in* B.[31]

Then, for agents $a \in B$: their perceived colouring at s_{new} is also inherited from $s_{last(\mathbf{M})}$.[32] Their perceived visibility at s_{new} is given, for every $\ell \in A$, as

$$s_{new}.\nu^a(\ell) := \begin{cases} X & \text{if } \ell = d \text{ and } d \notin s_{last}.\nu(a) \\ s_{last(\mathbf{M})}.\nu^a(\ell) \cup \{d\} & \text{if } \ell = a \\ s_{last(\mathbf{M})}.\nu^a(\ell) & \text{otherwise} \end{cases}$$

so a's perceived visibility about ℓ becomes X only if ℓ is the involved agent and she *could not* see it (equivalently, *cannot* see it, as factual visibility in s_{new} is exactly as in s_{last}), it is extended with d for a's self-perceived visibility, and remains as before otherwise.

Note how an agent's perceived self-visibility is not affected when she receives information about her own visibility. Indeed, suppose $d \in B$. In the formerly last state s_{last}, every agent can see herself, so $d \in s_{last}.\nu(d)$. Thus, the definition falls in the second case, yielding $s_{new}.\nu^d(d) = s_{last(\mathbf{M})}.\nu^d(d) \cup \{d\}$. But, again, d can already see herself, so this is simply $s_{last(\mathbf{M})}.\nu^d(d)$.

6.3.15. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$*,* $d \in A$*,* $X \subseteq A \cup O$ *with* $d \in X$*, and the structure* $\mathbf{M}_{B(d:=X)}$ *as in Definition 6.3.14. Then,* $\mathbf{M}_{B(d:=X)}$ *is a* MPTVM.

[30] More precisely, $s_{new}.\kappa(o) := s_{last(\mathbf{M})}.\kappa(o)$ for all $o \in O$, and $s_{new}.\nu(\ell) := s_{last(\mathbf{M})}.\nu(\ell)$ for every $\ell \in A$.

[31] Thus, for each $a \in A \setminus B$, we have $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for all $o \in O$, and $s_{new}.\nu^a(\ell) := s_{last(\mathbf{M})}.\nu^a(\ell)$ for every $\ell \in A$.

[32] That is, $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for every $o \in O$.

Proof:
Factual and perceived colouring in the new state s_{new} are inherited from $s_{last(\mathbf{M})}$, so objects have proper (factual and perceived) colour. To verify that, at s_{new}, every agent can factually and perceivably see herself, note that factual visibility is inherited from $s_{last(\mathbf{M})}$, and so is perceived visibility for agents not in B. Then, for agents in B, perceived visibility is inherited or extended in two of the three cases, the only exception being when $\ell = d$. But in this case, the new perceived visibility is X, which is required to contain d herself. Thus, in every visibility function in s_{new}, every agent can see herself. Finally, *perceived self-visibility* also holds at s_{new} because, while factual visibility is inherited, perceived visibility is either inherited or else extended in two of the three cases. The only case in which it is not is not relevant for this property, as we are discussing the $a = \ell$ scenario, and the further $\ell = d$ makes $d \notin s_{last}.\nu(a)$ impossible (recall: every agent can always see herself). □

The operation also preserves one of the optional requirements.

6.3.16. PROPOSITION. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$*,* $d \in A$*,* $X \subseteq A \cup O$ *with* $d \in X$*, and the structure* $\mathbf{M}_{B(d:=X)}$ *as in Definition 6.3.14. If* $\mathbf{M}$ *satisfies* correct perceived colour for observed objects*, then so does* $\mathbf{M}_{B(d:=X)}$.

Proof:
Immediate, as factual visibility in the new state s_{new} is exactly as in $s_{last(\mathbf{M})}$, and both factual and perceived colouring are inherited from $s_{last(\mathbf{M})}$. □

Still, the operation does not preserve *correct perceived visibility for observed agents*. Again, the action makes the agent believe she can see the involved agent, when this might actually not be the case. Since an agent can always see herself, she will not have a correct perception about her own visibility.

Acts of *private* communication The just defined actions of communication (Definition 6.3.11 and Definition 6.3.14) are not necessarily private. Take for example the action of communicating the colour of an object. As mentioned before, the agents receiving the information (those in the set B) make no assumption about whether anybody else received the information. As it will be shown later (Section 6.3.1), this is enough for modelling the Bake-Sale Task. Indeed, when the mother tells John about the product for sale, privacy is not necessary because, in John's perceived visibility, Mary couldn't see the object anyway.

There are, however, cases whereby the communication is clearly private. More importantly, this secrecy has consequences in the agents' perception of the situation. For example, consider a scenario in which two agents a and b can see each other, and yet none of them can see a given object. Suppose, moreover, that both think the object is white, when in reality it is black. If a is told privately about the object's real colour, then she will see it as black. But, additionally, the secrecy of the message should make her think that b cannot "see" the object (i.e. that b does not have up-to-date information about its colour). This is a case of communicated colour change whose private nature needs to be explicitly represented. To that end, we provide the following definition.

6.3.17. Definition (Privately communicating an object's colour). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Let $B \subseteq A$ be a set of agents and $p \in O$ an object, with $c \in R_p$ a proper colour.

The private communication $p{:=}c$ for agents in B produces the MPTVM $\mathbf{M}^{\mathsf{P}}_{B(p:=c)} = s_1 \cdots s_n s_{new}$. In the new multi-agent state s_{new}, factual functions are inherited from the last state in $\mathbf{M}$,[33] just like the perceived functions for agents *not in* B.[34]

Then, for agents $a \in B$: their perceived colouring at s_{new} is given, for every $o \in O$, as

$$s_{new}.\kappa^a(o) := \begin{cases} c & \text{if } o = p \text{ and } p \notin s_{last}.\nu(a) \\ s_{last(\mathbf{M})}.\kappa^a(o) & \text{otherwise} \end{cases}$$

so a's perceived colour about o becomes c only if this is the involved object and she *could not* see it, remaining as before otherwise. Then, the crucial difference with respect to Definition 6.3.11: the perceived visibility of every $a \in B$ at s_{new} is given, for every $\ell \in A$, as

$$s_{new}.\nu^a(\ell) := \begin{cases} s_{last(\mathbf{M})}.\nu^a(\ell) \cup \{p\} & \text{if } \ell = a \\ s_{last(\mathbf{M})}.\nu^a(\ell) \setminus \{p\} & \text{otherwise} \end{cases}$$

so a's perceived visibility about herself is extended with p, and her perceived visibility about all other agents is as before but now the object p is excluded.

This operation, a small variation of that for communicating the colour of an object, inherits the same behaviour with respect to the preservation of model properties.

6.3.18. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$, $p \in O$, $c \in R_p$ *and the structure* $\mathbf{M}^{\mathsf{P}}_{B(p:=c)}$ *as in Definition 6.3.17. Then,* $\mathbf{M}^{\mathsf{P}}_{B(p:=c)}$ *is a* MPTVM*.*

Proof:
For proper colouring, the argument is just as in Proposition 6.3.12. For self-visibility, every agent can see herself in the new state s_{new} under both factual

[33] More precisely, $s_{new}.\kappa(o) := s_{last(\mathbf{M})}.\kappa(o)$ for all $o \in O$, and $s_{new}.\nu(\ell) := s_{last(\mathbf{M})}.\nu(\ell)$ for every $\ell \in A$.

[34] Thus, for each $a \in A \setminus B$, we have $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for all $o \in O$, and $s_{new}.\nu^a(\ell) := s_{last(\mathbf{M})}.\nu^a(\ell)$ for every $\ell \in A$.

and perceived visibility: for the first because it is inherited, and for the second because in the only case in which it is reduced, the removed entity is an object. For *perceived self-visibility*, take any agent a. The property also holds at the new state because, while agent a's factual visibility $\nu(a)$ is inherited, her perceived visibility of *herself*, $\nu^a(a)$, is never reduced. □

6.3.19. PROPOSITION. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$*,* $p \in O$*,* $c \in R_p$ *and the structure* $\mathbf{M}^{\mathsf{P}}_{B(p:=c)}$ *as in Definition 6.3.17. If* $\mathbf{M}$ *satisfies* correct perceived colour for observed objects*, so does* $\mathbf{M}^{\mathsf{P}}_{B(p:=c)}$*.*

Proof:
As in the proof of Proposition 6.3.13. □

6.3.20. FACT. The operation for private communication of an object's colour does not preserve *correct perceived visibility for observed agents*. As in the non-private case, communication makes the agent believe she can see the involved object, when in fact this might not be the case. Since an agent can always see herself, she will not have a correct perception about her own visibility.

The analogous case, now for privately communicating visibility change:

6.3.21. DEFINITION (Privately communicating an agent's visibility). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM. Let $B \subseteq A$ be a set of agents and $d \in A$ an agent, with $X \subseteq A \cup O$ a set of agents and objects satisfying $d \in X$.

The private communication $d{\leftarrow}X$ for agents in B produces the MPTVM $\mathbf{M}^{\mathsf{P}}_{B(d\leftarrow X)} = s_1 \cdots s_n s_{new}$. In the new multi-agent state s_{new}, factual functions are inherited from the last state in $\mathbf{M}$,[35] just like the perceived functions for agents *not in* B.[36]

Then, for agents $a \in B$: their perceived colouring at s_{new} is also inherited from $s_{last(\mathbf{M})}$.[37] Then, the crucial difference with respect to Definition 6.3.14: their perceived visibility at s_{new} is given, for every $\ell \in A$, as

$$s_{new}.\nu^a(\ell) := \begin{cases} X & \text{if } \ell = d \text{ and } d \notin s_{last}.\nu(a) \\ s_{last(\mathbf{M})}.\nu^a(\ell) \cup \{d\} & \text{if } \ell = a \\ s_{last(\mathbf{M})}.\nu^a(\ell) \setminus \{d\} & \text{if } \ell \notin \{a, d\} \\ s_{last(\mathbf{M})}.\nu^a(\ell) & \text{if } \ell = d \text{ and } d \in s_{last}.\nu(a) \end{cases}$$

[35] More precisely, $s_{new}.\kappa(o) := s_{last(\mathbf{M})}.\kappa(o)$ for all $o \in O$, and $s_{new}.\nu(\ell) := s_{last(\mathbf{M})}.\nu(\ell)$ for every $\ell \in A$.

[36] Thus, for each $a \in A \setminus B$, we have $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for all $o \in O$, and $s_{new}.\nu^a(\ell) := s_{last(\mathbf{M})}.\nu^a(\ell)$ for every $\ell \in A$.

[37] That is, $s_{new}.\kappa^a(o) := s_{last(\mathbf{M})}.\kappa^a(o)$ for every $o \in O$.

so a's perceived visibility about ℓ becomes X only if this is the involved agent and she *could not* see it and it is extended with d for a's self-perceived visibility. For agents other than herself and the involved agent d, the latter is excluded from the previous perceived visibility, remaining as before otherwise.

Here is the behaviour of this operation, property-preservation-wise.

6.3.22. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$, $d \in A$, $X \subseteq A \cup O$ *with* $d \in X$*, and the structure* $\mathbf{M}^{\mathsf{P}}_{B(d:=X)}$ *as in Definition 6.3.21. Then,* $\mathbf{M}^{\mathsf{P}}_{B(d:=X)}$ *is a* MPTVM.

Proof:
For proper colouring, the argument is the same as in Proposition 6.3.15. For self-visibility, the only cases in which the property is at risk (i.e. not inherited) are those for the perception of an agent $a \in B$ about the visibility of any agent but her. If the agent is d, then the requirement on X guarantees $d \in s_{new}.\nu^a(d)$; if the agent ℓ is any other, from $\ell \in s_{last}.\nu^a(\ell)$ and $d \neq \ell$ it follows that $\ell \in s_{last}.\nu^a(\ell) \setminus \{d\}$, that is, $\ell \in s_{new}.\nu^a(\ell)$. Finally, *perceived self-visibility* (the $a = \ell$ scenario) also holds at s_{new} because, while factual visibility is inherited, perceived visibility is either inherited or else extended in two of the four cases. The only cases in which it is not are not relevant for this property: the requirement "$\ell \notin a, d$" puts us out of our relevant scenarios, and the requirement $\ell = d$ makes $d \notin s_{last}.\nu(a)$ impossible (every agent can see herself). □

6.3.23. Proposition. *Let* $\mathbf{M}$ *be a* MPTVM*; take* $B \subseteq A$, $d \in A$, $X \subseteq A \cup O$ *with* $d \in X$*, and the structure* $\mathbf{M}^{\mathsf{P}}_{B(d:=X)}$ *as in Definition 6.3.14. If* $\mathbf{M}$ *satisfies* correct perceived colour for observed objects*, then so does* $\mathbf{M}^{\mathsf{P}}_{B(d:=X)}$.

Proof:
Exactly as in Proposition 6.3.16. □

6.3.24. Fact. The operation for private communication of an agent's visibility does not preserve *correct perceived visibility for observed agents*. As in the non-private case, communication makes the agent believe she can see the involved agent, when in fact this might not be the case. Since an agent can always see herself, she will not have a correct perception about her own visibility.

A formal language The language is extended to include communication actions.

6.3.25. Definition (Language $\mathcal{L}_{\mathscr{C}}$). The formulas of the language $\mathcal{L}_{\mathscr{C}}$ are given in the lines of Definition 6.2.6; the only difference is in the construction of α, which now becomes

$$\alpha ::= p_1{:=}c_1, \ldots, p_k{:=}c_k \mid b_1{\leftarrow}X_1, \ldots, b_h{\leftarrow}X_h \mid B(p{:=}c) \mid B(d{\leftarrow}X) \mid B^{\mathsf{P}}(p{:=}c) \mid B^{\mathsf{P}}(d{\leftarrow}X)$$

for $k \geqslant 1$, $\{p_1, \ldots, p_k\} \subseteq O$, $c_i \in R_{p_i}$, $h \geqslant 1$, $\{b_1, \ldots, b_h\} \subseteq A$, $X_i \subseteq A \cup O$, $b_i \in X_i$, $B \subseteq A$, $p \in O$, and $c \in R_p$, $d \in A$, $X \subseteq A \cup O$.

Here is, then, the semantic interpretation of formulas in the new language under the new structures. The crucial change is in the interpretation of mental attribution formulas, whose visibility and colouring condition become relativized to the perception of the main agent in the attribution.

6.3.26. Definition (Semantic interpretation). Let $\mathbf{M} = s_1 \cdots s_n$ be a MPTVM.

- Take $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c)$. Its a_1-*relativized visibility condition* on a state s, denoted by $\mathrm{vis}^{a_1}_{\chi}(s)$, is given by

$$\mathrm{vis}^{a_1}_{\chi}(s) \quad \textit{iff}_{\,def} \quad a_2 \in s.\nu^{a_1}(a_1) \;\&\; \ldots \;\&\; a_k \in s.\nu^{a_1}(a_{k-1}) \;\&\; o \in s.\nu^{a_1}(a_m)$$

Then,

$$M \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c) \quad \textit{iff}_{\,def} \quad \mathop{\rotatebox[origin=c]{180}{\Large\&}}_{i=0}^{n-1} \left(\begin{array}{c} \overbrace{\mathrm{vis}^{a_1}_{\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o\lhd c)}(s_{n-i})}^{\mathsf{vis}} \;\&\; \overbrace{s_{n-i}.\kappa^{a_1}(o) = c}^{\mathsf{col}} \\ \& \\ \underbrace{\mathop{\Large\&}_{j=1}^{i} \text{not } \mathrm{vis}^{a_1}_{\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o\lhd c)}(s_{n-(j-1)})}_{\mathsf{no-latter-vis}} \end{array} \right)$$

Boolean operators are interpreted as usual. For action modalities,

$$\mathbf{M} \Vdash [\alpha]\,\phi \qquad \textit{iff}_{\,def} \qquad \mathbf{M}_{[\alpha]} \Vdash \phi$$

The only change in the evaluation of mental attribution formulas is the relativization of both their visibility and their colouring conditions. Thus, one can express their semantic interpretation in a different but equivalent way. Indeed, one can use the formulas' original semantic interpretation (Definition 6.2.7), not in the multi-perspective model, but rather in a suitable relativized one.

6.3.27. Definition (Agent-relativized state and model).

- Take an agent $a_1 \in A$ and a multi-agent state $s = \langle \kappa, \nu, \{\kappa^a, \nu^a\}_{a \in A} \rangle$. Its a_1-relativized (simple) state is defined as $s|_{a_1} := \langle \kappa^{a_1}, \nu^{a_1} \rangle$.
- Take a MPTVM $\mathbf{M} = s_1 \ldots s_n$. Its a_1-relativized TVM is defined as $\mathbf{M}|_{a_1} := s_1|_{a_1} \cdots s_n|_{a_1}$.

6.3.28. Observation. For every MPTVM $\mathbf{M}$ and every attribution formula $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c)$,

$$\mathbf{M} \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_k}(o{\lhd}c) \quad \text{iff} \quad \mathbf{M}|_{a_1} \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_k}(o{\lhd}c).$$

6.3.1 Modelling FBTs

Here is how the multi-perspective TV framework can be used for representing FBTs involving communication.

6.3.29. Example (Second-order FBT: the Bake-Sale Task, continued). We break down the Bake-Sale Task scenario:

1. Mary and John hear that the church is having a bake sale, where chocolate cookies are sold.
2. Mary goes to the church to get chocolate cookies.
3. Mom comes back home and tells John that according to her, they only sell pumpkin pie.
4. Mary arrives at the bake sale but all there is for sale is brownies.
5. At some point, John knows that Mary must have arrived at the bake sale.

Let's spell out the depiction of the task in model **M** on Figure 6.3. **(s0)** Initially (Example 6.3.4), the product is brownies (the object is white), and both Mary and John can only see each other. However, they both think they can see everything and that the product is cookies (the object is black). **(s1)** Afterwards, Mary leaves for the church. This amounts to a visibility change for both Mary and John, as they stop seeing each other. This also has an effect on their perceptions. **(s2)** Then, John is told that the product is pumpkin pie (the object is gray). This is captured by an action of communication of colour, which entails both that the product is in John's perceived visibility and that he thinks it is pumpkin pie. Still, nothing changes in the real situation and in Mary's perception. **(s3)** Mary arrives at the bake sale and sees that the product is brownies (the object is white). This is a visibility change for Mary, making her see the product as it truly is, and affecting her perception too. **(s4)** Eventually, John knows that Mary must have arrived at the bake sale. This is captured by a communication of visibility change. While nothing changes in the real situation and in Mary's perception, John now thinks that Mary can see the product.

According to the model:

- $\mathbf{M} \Vdash \mathrm{B}_M(p \triangleleft \mathit{white}) \wedge \mathrm{B}_J(p \triangleleft \mathit{gray})$, i.e. Mary believes that the product is brownies (because she can actually see it) while John believes that the product is pie (because of the false information given to him).
- $\mathbf{M} \Vdash \mathrm{B}_M \mathrm{B}_J(p \triangleleft \mathit{black}) \wedge \mathrm{B}_J \mathrm{B}_M(p \triangleleft \mathit{gray})$, i.e. Mary believes that John believes that the product is cookies (because of backtracking to s_0), while John believes that Mary believes that the product is pie (because of his perceived situation in s_4).
- $\mathbf{M} \Vdash \mathrm{B}_M \mathrm{B}_J \mathrm{B}_M(p \triangleleft \mathit{black}) \wedge \mathrm{B}_J \mathrm{B}_M \mathrm{B}_J(p \triangleleft \mathit{black})$, i.e. Mary believes that John believes that Mary believes that the product is cookies, and John believes that Mary believes that John believes that the product is cookies (because of backtracking to s_0).

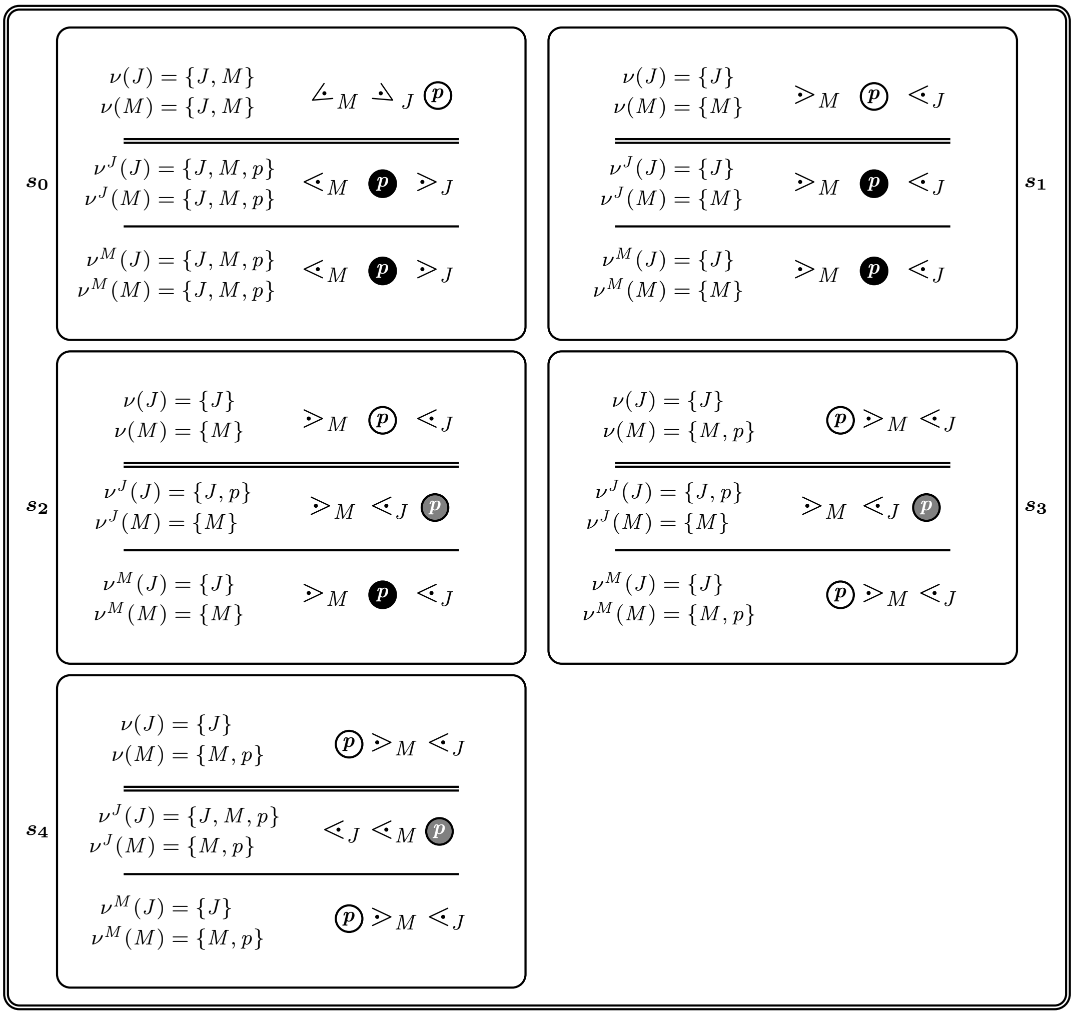


Figure 6.3: MPTVM for the Bake-Sale Task.

As mentioned above, the Bake-Sale Task does not require private communication. However, the Ice-cream Truck Task described below does.

6.3.30. Example (Second-order FBT (the *Ice-cream Truck* Task)). The task, adapted from Perner and Wimmer (1985), can be seen as the following sequence.

1. John and Mary are in the park where they see an ice-cream truck.
2. Mary leaves to go home and get money for an ice-cream.
3. The ice-cream truck driver tells John that he will go to the church.
4. The ice-cream truck is on its way to the church.
5. Mary sees the truck from her home's window. She comes out and asks the driver where he's heading to (namely, the church).
6. The truck arrives at the church.

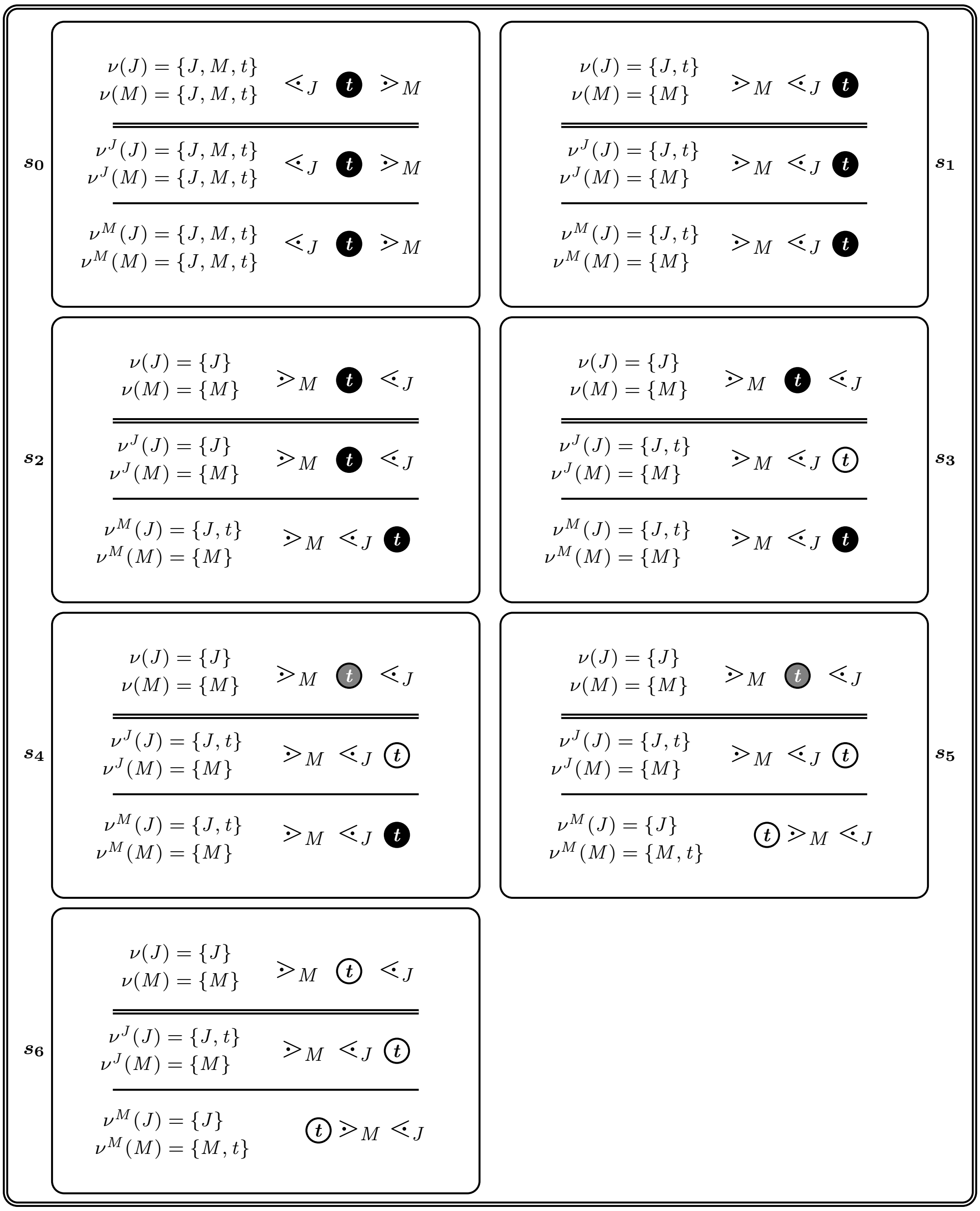


Figure 6.4: MPTVM for the Ice-cream Truck Task.

Let's spell out the modelling of the task, as depicted in Figure 6.4. **(s0)** Initially, both Mary and John can see each other and the truck being in the park (the object being black). **(s1)** Then Mary leaves. This amounts to a visibility change for both, as afterwards Mary only sees herself, and John only sees himself and the truck. This also affects their perceptions. **(s2)** John stops seeing the truck in the park. This is a visibility change, affecting the real situation and John's perception. **(s3)** But John has been told that the truck will go to the church (the object will be white). This is a communication of colour for John.[38] **(s4)** The truck leaves the park and is on the road to the church (i.e. it becomes gray). This is a change of colour, witnessed neither by John nor by Mary. **(s5)** Mary talks with the truck driver and gets private information about the truck's location. Now she thinks the truck will be in the church (the object will be white), and she does not think that John has proper information about the truck anymore. This is a private communication change for Mary. **(s6)** Finally, the truck arrives at the church. This is a colour change.

As a result:

- $\mathbf{M} \Vdash \mathrm{B}_M(t \lhd \mathit{white}) \wedge \mathrm{B}_J(t \lhd \mathit{white})$, i.e. both agents believe that the truck is in the church, because they have been told so.
- $\mathbf{M} \Vdash \mathrm{B}_M \mathrm{B}_J(t \lhd \mathit{black}) \wedge \mathrm{B}_J \mathrm{B}_M(t \lhd \mathit{black})$, i.e. Mary believes that John believes that the truck is in the park, and John believes that Mary believes that the truck is in the park (because of backtracking to s_0), and likewise for
- $\mathbf{M} \Vdash \mathrm{B}_M \mathrm{B}_J \mathrm{B}_M(t \lhd \mathit{black}) \wedge \mathrm{B}_J \mathrm{B}_M \mathrm{B}_J(t \lhd \mathit{black})$

6.3.2 Some technical results

Section 6.2.2 presented some technical results concerning the basic TV framework. This section discusses the same issues, modal translation and bisimulation, now for the full MPTVM setting.

Modal translation. The discussed connection between the TV setting and a more standard linear temporal framework, making crucial use of the strict *Since* operator, clarified what evaluating mental attributions amounts to. A connection can be also established between these linear temporal structures and the MPTV setting. The most straightforward way of doing so is taking each multi-agent state in a MPTVM to be a world in a linear temporal structure, just like in Definition 6.2.11. However, a multi-agent state contains not only factual visibility and factual colouring: it also contains the visibility and colouring as seen from the perspective of the different agents. In order to account for that, additional atomic propositions are needed. Here we focus on the action-less fragment of the MPTV setting; actions can be dealt with as in Section 6.2.2.

6.3.31. Definition (Multi-perspective derived linear structure). Let $\mathbf{M}$ be a MPTVM with A the set of agents, O the set of objects, and R_o a set of possible

[38] In the original story, John is merely told about the new location of the truck. We break this into two steps: one whereby John stops seeing the truck (**s2**) to capture that John will cease to see the truck, and one whereby John is informed about the new location (**s3**).

values for each object $o \in O$. Define the set of atoms

$$P^{\mathscr{C}}_{A,O,R_o} := \bigcup \left\{ \begin{array}{l} \{\lhd_a x \mid a \in A \text{ and } x \in A \cup O\}, \\ \{o{\lhd}c \mid o \in O \text{ and } c \in R_o\}, \\ \{b{:}\lhd_a x \mid b, a \in A \text{ and } x \in A \cup O\}, \\ \{b{:}o{\lhd}c \mid b \in A, o \in O \text{ and } c \in R_o\} \end{array} \right\}$$

The linear structure $\mathsf{M}^{\mathscr{C}}_{\mathbf{M}} = \langle W_{\mathbf{M}}, \prec_{\mathbf{M}}, V^{\mathscr{C}}_{\mathbf{M}} \rangle$ over $P^{\mathscr{C}}_{A,O,R_o}$ extends the linear structure of Definition 6.2.11 over P_{A,O,R_o} by allowing additional agent-specific atoms. More precisely, while domain $W_{\mathbf{M}}$ and temporal relation $\prec_{\mathbf{M}}$ mimic the finite sequence of states in $\mathbf{M}$ (see Definition 6.2.11), the valuation function $V^{\mathscr{C}}_{\mathbf{M}}$ is given by

$$\begin{array}{rr} V^{\mathscr{C}}_{\mathbf{M}}(\lhd_a x) := \{w_s \in W_{\mathbf{M}} \mid x \in s.\nu(a)\}, & V^{\mathscr{C}}_{\mathbf{M}}(b{:}\lhd_a x) := \{w_s \in W_{\mathbf{M}} \mid x \in s.\nu^b(a)\}, \\ V^{\mathscr{C}}_{\mathbf{M}}(o{\lhd}c) := \{w_s \in W_{\mathbf{M}} \mid s.\kappa(o) = c\}, & V^{\mathscr{C}}_{\mathbf{M}}(b{:}o{\lhd}c) := \{w_s \in W_{\mathbf{M}} \mid s.\kappa^b(o) = c\}. \end{array}$$

On the syntactic side,

6.3.32. Definition. Formulas of the modal language $\mathcal{L}^+_{\mathrm{S}}$ over P^{C}_{A,O,R_o} are given by:

$$\phi ::= \lhd_a x \mid o{\lhd}c \mid b{:}\lhd_a x \mid b{:}o{\lhd}c \mid \neg\phi \mid \phi \wedge \phi \mid \mathrm{S}(\phi, \phi)$$

for $a \in A$, $x \in A \cup O$ and $c \in R_o$. For their semantic interpretation over *pointed* temporal structures, atoms $\lhd_a x$, $o{\lhd}c$, $b{:}\lhd_a x$ and $b{:}o{\lhd}c$ are interpreted in the natural way, Boolean operators are interpreted as usual, and the *since* modality is interpreted as in Definition 6.2.10

Finally, here is the correspondence.

6.3.33. Proposition. *Given a mental attribution formula* $\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c)$ *in* $\mathcal{L}'$*, define the* $\mathcal{L}^+_{\mathrm{S}}$*-formula*

$$\mathrm{vis}^{a_1}_{a_1 \cdots a_m o} := a_1{:}\lhd_{a_1} a_2 \;\wedge\; \cdots \;\wedge\; a_1{:}\lhd_{a_{m-1}} a_m \;\wedge\; a_1{:}\lhd_{a_m} o,$$

expressing the visibility condition of the mental attribution formula from the main agent's (a_1's) perspective. Define the translation $tr^+ : \mathcal{L}' \to \mathcal{L}^+_{\mathrm{S}}$ *as*

$$\begin{aligned} tr^+(\mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o{\lhd}c)) &:= \bigvee \left\{ \begin{array}{l} \mathrm{vis}^{a_1}_{a_1 \cdots a_m o} \;\wedge\; a_1{:}o{\lhd}c, \\ \neg\, \mathrm{vis}^{a_1}_{a_1 \cdots a_m o} \;\wedge\; \mathrm{S}(\mathrm{vis}^{a_1}_{a_1 \cdots a_m o} \wedge a_1{:}o{\lhd}c, \neg\, \mathrm{vis}^{a_1}_{a_1 \cdots a_m o}) \end{array} \right\}, \\ tr^+(\neg\phi) &:= \neg tr^+(\phi), \\ tr^+(\phi \wedge \psi) &:= tr^+(\phi) \wedge tr^+(\psi). \end{aligned}$$

Then, for any MPTVM $\mathbf{M}$ *and any* $\phi \in \mathcal{L}'$*,*

$$\mathbf{M} \Vdash \phi \qquad \textit{iff} \qquad (\mathsf{M}^{\mathscr{C}}_{\mathbf{M}}, w_{s_{last(\mathbf{M})}}) \Vdash tr^{+}(\phi).$$

Proof:
The proof is by induction on $\mathcal{L}'$. For the crucial base case, for mental attribution formulas, recall that $\mathbf{M} \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_k}(o \lhd c)$ iff $\mathbf{M}|_{a_1} \Vdash \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_k}(o \lhd c)$ (by Observation 6.3.28). Then, by Proposition 6.2.13, the latter is equivalent to $(\mathsf{M}_{\mathbf{M}|_{a_1}}, w_{s_{last(\mathbf{M}|_{a_1})}}) \Vdash tr(\phi)$. For the final step, note how $(\mathsf{M}_{\mathbf{M}|_{a_1}}, w_{s_{last(\mathbf{M}|_{a_1})}}) \Vdash tr(\phi)$ evaluates the non-relativized formula $tr(\phi)$ in the relativized pointed model $(\mathsf{M}_{\mathbf{M}|_{a_1}}, w_{s_{last(\mathbf{M}|_{a_1})}})$. This is equivalent to evaluating the relativized formula $tr^{+}(\phi)$ in a non-relativized pointed model $(\mathsf{M}^{\mathscr{C}}_{\mathbf{M}}, w_{s_{last(\mathbf{M})}})$. □

There is a more "epistemic" way of connecting MPTVM's with more standard modal structures. The idea is that each multi-agent state in the MPTVM gives rise not to a *single* world, but rather to a *collection* of them: ***(i)*** the "factual" one, where the valuation is determined by the factual functions, and ***(ii)*** one for each agent, where the valuation is determined by the agent's perceived visibility and colouring functions. Then, besides the temporal relation between the "factual" worlds, the resulting relational structure uses agent-specific "epistemic" relations R_a, connecting each "factual" world to the one representing a's perspective. In this way, the information contained in the modal worlds of the structure of Definition 6.3.31 is split into different "simpler" worlds, with each agent's perception tracked through the epistemic relations.

A language for describing these *epistemic temporal* structures requires not only the *Since* operator, but also a belief modal operators B_a for each agent $a \in A$, which is semantically evaluated as normal modal operator over its corresponding relation R_a. Then, a translation from $\mathcal{L}'$ to this epistemic temporal language simply tweaks the formula expressing the visibility condition, using the belief operator B_a for the relativization.

Still, the technical details of this correspondence are relatively simple; thus, they will not be discussed here.

Bisimulation Much like the translation, the notion of the bisimulation, constructed in Definition 6.2.14, can be adapted to accommodate MPTVMs too.

6.3.34. Definition (MPTV-bisimilarity). A MPTV-*bisimulation* is a non-empty relation $Z \subseteq \mathbf{MPTVM} \times \mathbf{MPTVM}$ such that, if $\mathbf{M} = s_1 \cdots s_n$ and $\mathbf{M}' = s'_1 \cdots s'_{n'}$ are such that $(\mathbf{M}, \mathbf{M}') \in Z$, then the following holds.

(atom) For every mental attribution formula $\chi := \mathrm{B}_{a_1} \cdots \mathrm{B}_{a_m}(o \lhd c)$, there is $s_i \in \mathbf{M}$ such that ***(i)*** $\mathrm{vis}^{a_1}_{\chi}(s_i)$ holds, ***(ii)*** $s_i.\kappa^{a_1}(o) = c$ and ***(iii)*** $\mathrm{vis}^{a_1}_{\chi}(s_k)$ fails for every $s_k \in \mathbf{M}$ with $s_i < s_k$ iff there is $s'_j \in \mathbf{M}'$ such that ***(i)*** $\mathrm{vis}^{a_1}_{\chi}(s'_j)$ holds, ***(ii)*** $s'_j.\kappa^{a_1}(o) = c$ and ***(iii)*** $\mathrm{vis}^{a_1}_{\chi}(s'_h)$ fails for every $s'_h \in \mathbf{M}'$ with $s'_j < s'_h$.

Two MPTVMs $\mathbf{M}$ and $\mathbf{M}'$ are said to be MPTV-*bisimilar* (notation: $\mathbf{M} \mathrel{\underline{\leftrightarrow}} \mathbf{M}'$) iff there is a MPTV-bisimulation Z with $(\mathbf{M}, \mathbf{M}') \in Z$.

Again, if two $\mathbf{M}$ and $\mathbf{M}'$ are bisimilar, then they satisfy the same static formulas. This definition does not preserve full equivalence, in particular for action formulas, for the same reason as before.

6.3.35. Proposition. *For any two* MPTVM*s* $\mathbf{M}$ *and* $\mathbf{M}'$,

$$\mathbf{M} \mathrel{\underline{\leftrightarrow}} \mathbf{M}' \qquad \textit{iff} \qquad \mathbf{M} \leftrightsquigarrow_{\mathcal{L}'} \mathbf{M}'$$

Proof:
It follows from the requirements of **atom**. □

We can also extend the dynamic bisimulation case (Section 6.2.2) to the framework with communication. First, fix $\mathbf{MPTVM}$ to denote the class of all MPTVMs over the fixed sets A, O and $\{R_o \mid o \in O\}$.

6.3.36. Definition (R-based MPTV-bisimilarity). Let $\mathbf{R} = \{\mathsf{R}_\iota \subseteq \mathbf{MPTVM} \times \mathbf{MPTVM} \mid \iota \in I\}$ be a family of relations between MPTVMs. A $\mathbf{R}$-*based* MPTV-*bisimulation* is a non-empty relation $Z \subseteq \mathbf{MPTVM} \times \mathbf{MPTVM}$ such that, for every $\mathbf{M} = s_1 \cdots s_n$ and $\mathbf{M}' = s'_1 \cdots s'_{n'}$ satisfying $(\mathbf{M}, \mathbf{M}') \in Z$, the following statements hold.

(atom) As in Definition 6.3.34.

(R-transition) For every $\iota \in I$,

> **(ι-forth)** If there is $\mathbf{N} \in \mathbf{MPTVM}$ such that $(\mathbf{M}, \mathbf{N}) \in \mathsf{R}_\iota$ for some $\iota \in I$, then there is $\mathbf{N}' \in \mathbf{MPTVM}$ such that $(\mathbf{M}', \mathbf{N}') \in \mathsf{R}_\iota$ and $(\mathbf{N}, \mathbf{N}') \in Z$.
>
> **(ι-back)** Vice versa.

Two MPTVMs $\mathbf{M}$ and $\mathbf{M}'$ are said to be $\mathbf{R}$-MPTV-*bisimilar* (notation: $\mathbf{M} \mathrel{\underline{\leftrightarrow}}_{\mathbf{R}} \mathbf{M}'$) iff there is a $\mathbf{R}$-based MPTV-bisimulation Z with $(\mathbf{M}, \mathbf{M}') \in Z$.

The transitions that matter in the multi-perspective framework are the functional ones that take a MPTVM to one that results from ***(i)*** factual changes, both on colour and visibility, ***(ii)*** communication, both "unbiased" (i.e. no assumption about who else received the information) and private, both about colour and visibility.

6.3.37. Definition (Family $\mathbf{R}^{\mathscr{C}}$). Recall that $\mathtt{C}$ and $\mathtt{V}$ are, respectively, the set of all proper colour and visibility assignments over A, O and $\{R_o \mid o \in O\}$.

- For each $C \in \mathtt{C}$, the relation $\mathsf{R}_C \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}_C := \{(\mathbf{M}, \mathbf{M}_C) \mid \mathbf{M} \in \mathbf{MPTVM}\}$.

- For each $V \in \mathtt{V}$, the relation $\mathsf{R}_V \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}_V := \{(\mathbf{M}, \mathbf{M}_V) \mid \mathbf{M} \in \mathbf{MPTVM}\}$.

Then, let $\mathtt{C}^{\mathscr{C}}$ and $\mathtt{V}^{\mathscr{C}}$ be, respectively, the sets of proper messages communicating colour and visibility over A, O and $\{R_o \mid o \in O\}$.

- For each $C^{\mathscr{C}} \in \mathtt{C}^{\mathscr{C}}$, the relation $\mathsf{R}_{C^{\mathscr{C}}} \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}_{C^{\mathscr{C}}} := \{(\mathbf{M}, \mathbf{M}_{C^{\mathscr{C}}}) \mid \mathbf{M} \in \mathbf{MPTVM}\}$, and the relation $\mathsf{R}^{\mathsf{P}}_{C^{\mathscr{C}}} \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}^{\mathsf{P}}_{C^{\mathscr{C}}} := \{(\mathbf{M}, \mathbf{M}^{\mathsf{P}}_{C^{\mathscr{C}}}) \mid \mathbf{M} \in \mathbf{MPTVM}\}$.
- For each $V^{\mathscr{C}} \in \mathtt{V}^{\mathscr{C}}$, the relation $\mathsf{R}_{V^{\mathscr{C}}} \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}_{V^{\mathscr{C}}} := \{(\mathbf{M}, \mathbf{M}_{V^{\mathscr{C}}}) \mid \mathbf{M} \in \mathbf{MPTVM}\}$, and the relation $\mathsf{R}^{\mathsf{P}}_{V^{\mathscr{C}}} \subseteq (\mathbf{MPTVM} \times \mathbf{MPTVM})$ is given by $\mathsf{R}^{\mathsf{P}}_{V^{\mathscr{C}}} := \{(\mathbf{M}, \mathbf{M}^{\mathsf{P}}_{V^{\mathscr{C}}}) \mid \mathbf{M} \in \mathbf{MPTVM}\}$.

The family of relations $\mathbf{R}^{\mathscr{C}}$ is defined as

$$\mathbf{R}^{\mathscr{C}} := \bigcup \left\{ \begin{array}{l} \{\mathsf{R}_C \mid C \in \mathtt{C}\}, \\ \{\mathsf{R}_V \mid V \in \mathtt{V}\}, \\ \{\mathsf{R}_{C^{\mathscr{C}}}, \mathsf{R}^{\mathsf{P}}_{C^{\mathscr{C}}} \mid C^{\mathscr{C}} \in \mathtt{C}^{\mathscr{C}}\}, \\ \{\mathsf{R}_{V^{\mathscr{C}}}, \mathsf{R}^{\mathsf{P}}_{V^{\mathscr{C}}} \mid V^{\mathscr{C}} \in \mathtt{C}^{\mathscr{C}}\} \end{array} \right\}.$$

6.3.38. Proposition. *For any two* MPTVM*s* $\mathbf{M}$ *and* $\mathbf{M}'$*,*

$$\mathbf{M} \underline{\leftrightarrow}_{\mathbf{R}^{\mathscr{C}}} \mathbf{M}' \qquad \textit{iff} \qquad \mathbf{M} \leftrightsquigarrow_{\mathcal{L}_{\mathscr{C}}} \mathbf{M}'$$

Proof:
The proof works like the one of Proposition 6.2.20, adapted from TVMs to MPTVMs, from $\mathbf{R}_{\mathtt{C},\mathtt{V}}$ to $\mathbf{R}^{\mathscr{C}}$, and from $\mathcal{L}$ to $\mathcal{L}_{\mathscr{C}}$.

□

6.3.3 Evaluating the framework

The MPTV framework still follows the rationale of a backtracking process, but now taking into account the values of colour and visibility as seen from the perspective of the attributing agent. Thanks to this relatively simple change, it is possible to represent acts of communication.

It is interesting to notice how communication is another possible source of false belief attributions. On the one hand, false belief attributions arise in the TV framework only when the visibility condition fails at the last available state. In such cases, agents have to look back in the past for a state in which the needed information was available. The fact that the situation might have changed ever since is what leads them to attributing false beliefs. On the other hand, in the

MPTV framework, agents may attribute false beliefs also because of misleading information, as evinced by Examples 6.3.29 and 6.3.30.

Still, probably the most interesting thing to notice is the constant increase in complexity for operations representing actions. Indeed, factual changes in both colouring and visibility in the TV setting are straightforward: simply add a new state in which the colouring/visibility of the objects/agents being affected is updated appropriately (Definitions 6.2.3 and 6.2.4). In the MPTV setting, operations for factual change become slightly more complex, but not dramatically so: they only take the additional care of guaranteeing that agents observing the affected objects/agents will notice the change (Definitions 6.3.5 and 6.3.8). Things escalate with operations representing communication. The "unbiased" case needs to decide, additionally, what gets priority: visibility or communication (Definitions 6.3.11 and 6.3.14). Then, some scenarios (e.g. Example 6.3.30) might require to distinguish agents who receive the message from those who do not. This affects not only the way the receivers envision the involved object/agent, but also the way they envision both agents "in the loop" and agents "out of the loop" (Definitions 6.3.17 and 6.3.21). This is expected to become even more convoluted for types of communication that indicate not only who receives the message and who does not, but also who sees who receives the message, and who does not.

Compare this increase in complexity with the uniformity and relative simplicity of the operations representing the same actions (factual change and diverse [public, private, secret] forms of communication) in the DEL framework.[39] Those two types of changes are dealt with by essentially the same structures (action models plus postconditions), which can deal with further convoluted scenarios. This, which could be seen as a drawback of this proposal (e.g. from an implementation point of view), can be actually seen as a proof that the framework does a good work capturing the way humans keep track of information and make use of it. As it has been discussed, adults find it difficult to master tasks that require higher orders of ToM. The increased complexity of the operations that the models require for representing increasingly complex social scenarios might suggest that the setting is on the right track.

6.4 Conclusions and further questions

This chapter has introduced a temporal framework suitable for capturing how "real" agents attribute mental states on basis of visibility, memory, and communication, inspired by well-known FBTs. Contrary to EL, this framework is built on a simple semantic model and a complex clause for interpreting mental state attributions. In its basic version (TV modelling) it addresses attributions of belief grounded on visibility and memory, while its extended version (MPTV modelling) further addresses the effect of communication on belief attributions.

[39] See, e.g. van Benthem et al. (2006); van Ditmarsch and Kooi (2008) for the first, and Baltag et al. (1998) for the second.

We have studied the technical features of both and implemented them on various FBTs, evaluating them against criteria of robustness and faithfulness. We have furthermore drawn more connections between this type of modelling and features of social cognition, and compared this proposal to others with overlapping goals or logical machinery. In the next chapter, we will cover another dimension of higher-order mental states: the rule-based manipulations of formed attributions.

The current project itself presents several lines of further research. Among the ones regarding extensions of the proposed framework, there are three relevant ones. First, the language used for describing MPTVM is enough for reasoning about belief attributions, their Boolean combinations, and the way they are affected by actions of factual change and communication. But one might be also interested in more expressive languages. One possibility is a language similar to that in EL, allowing the free use of epistemic operators as in formulas of the form $\mathrm{B}_M(\mathit{cho}\lhd\mathit{green} \wedge \mathrm{B}_J(\mathit{cho}\lhd\mathit{white}))$. On the side of the semantics, this might require shifting evaluation across multiple states, as different visibility conditions may be satisfied at different moments.

A second direction is towards modelling further types of FBTs. An interesting case is about tasks focusing on attributions to oneself. For example, in the *Smarties Task* (Gopnik and Astington, 1988), the question concerns an agent's belief about her own belief at an earlier point of time. The temporal nature of this framework may assist the study of these tasks as well, provided that suitable temporally indexed belief operators are introduced.

Then, a further goal is modelling more "suspicious" agents. Indeed, as mentioned in Section 6.2.3, while agents in the MPTV framework assume the current situation is exactly as the last time they had information about it (via visibility or communication), more sceptical ones might be careful about what has happened in the meantime. Modelling these agents asks for variations on the inertia assumption, which implies either changes on the visibility condition, or else its replacement by a more general *evaluation* requirement.

So far we have mentioned different types of attribution failures that can be derived from our semantics. These are in fact experienced at different points in one's life. At the beginning (<4 years of age), we tend to attribute our own beliefs to others, collapsing their perspective to our own, but we gradually master higher orders of ToM. Still, we remain subjects to cognitive limitations, situational constraints, misinformation and lack of visibility. With these ingredients in place, an interesting (albeit more general) project has to do with modelling the development of ToM in the course of our lives and our ability to learn how to use it. This topic has been addressed by Arslan et al. (2013), using the ACT-R cognitive architecture, to study the developmental transitions from zero- to second- order false belief reasoning.

The chapter has focused on the formation of attributions in FBTs. However, this is not the only type of tasks testing people's ToM. Meijering (2014) suggests the Marble Drop Games as alternatives to the use of FBT storylines, for they are more flexible and can be re-used on the same agents. Such games are also interesting to study because they reveal the *strategies* of agents, which involve not only having, in principle, the ability to apply ToM, but also the choice of the agent to do so. For example, following a more economical strategy (in accord with the features discussed in Section 6.2.3) with respect to ToM is validated by the computational model of Meijering et al. (2014), also discussed in Section 2.3.3. According to it, people might prefer simple, less cognitively costly strategies and, only when necessary, resort to strategies that require complex attributions. The connection of one type of storylines with another and the strategizing of agents, crucially depending on the cognitive costs one is willing to pay, is also a natural direction for future work.

Chapter 7

Inference, introspection, attribution

In this chapter, we continue our treatment of multi-agent reasoning, now focusing on the *manipulation* of beliefs.[1] This encompasses all three cases of reasoning delineated before: (a) deductive inference, (b) introspection, and (c) reasoning about the reasoning of others. While we have addressed deductive inference, we have not yet discussed the manipulations underlying (b) and (c). We now propose a unified approach, wherein agents come to believe what can be feasibly reached from their current belief state, with respect to all three cases of non-ideal reasoning. As in Part II, descriptive facts, e.g. on limitations of time and memory, are still instrumental to ensure that *ought* implies *can*, in accord with Part I.

More concretely, we model reasoning in a multi-agent environment as dynamic, costly, and rule-based. Dynamic, because beliefs, including higher-order ones, also evolve through actions of "internal elucidation" (van Benthem, 2008c). Costly, because experience as well as experimental results indicate that reasoning steps require effort and it is precisely because of that, and not because of an arbitrary bound, that reasoning processes eventually halt. These steps are viewed as instances of rule-based reasoning. Concerning deduction, each step is taken to correspond to an application of an inference rule (as in Part II). Concerning introspection, every nesting of belief is seen as a reasoning step, and in analogy to inference, we refer to it as an application of an introspective rule. Multi-agent interaction can be extremely diverse, involving more than reasoning alone, e.g. communication and observation. However, no matter how certain beliefs about others were formed, we here focus on how agents derive further information by attributing reasoning steps to others based on this initial stock.

The remainder of the chapter is organized as follows. Section 7.1 introduces some background notions that we rely on to design our framework. The framework to model bounded inference, introspection, and attribution, is in turn presented, along with examples, in Section 7.2. The implications of the framework and comparisons with other works are discussed in Section 7.3. We then devise a method that allows us to extract a sound and complete logic (Section 7.4), followed by some extensions of our framework (Section 7.5) and the conclusions (Section 7.6).

[1] The chapter is based on Solaki (2021b).

7.1 Background

The main ingredients for our framework are *action models* and *impossible-worlds semantics*. The background behind the latter has been already discussed (Section 3.1.1); in what follows, we give an overview of the former which will serve as the basis to provide a unified treatment of the various reasoning steps in a multi-agent environment. In order to do so, we first have to introduce the standard multi-agent (doxastic) Kripke models. Given a non-empty set of agents Ag and a set of propositional atoms Φ:

7.1.1. Definition (Multi-agent model). A *multi-agent model* M is a tuple $\langle W, \{\longrightarrow_j\}_{j \in Ag}, V\rangle$ where:

- W is a non-empty set of possible worlds.
- each $\longrightarrow_j$ is a binary relation on W, capturing the (doxastic) accessibility of agent j.
- V is a valuation which assigns to each world a set of propositional atoms from Φ (those that are true there).

The simple multi-agent doxastic language $\mathcal{L}_B$ is built by extending the propositional language with belief operators (B_j).

7.1.2. Definition (Multi-agent doxastic language). The language $\mathcal{L}_B$ of multi-agent doxastic logic is given by:

$$\phi \quad ::= \quad p \quad | \quad \neg\phi \quad | \quad \phi \wedge \phi \quad | \quad B_j\phi$$

with $p \in \Phi$.

Formulas are evaluated at pointed models. The Boolean clauses are treated in the usual way. The crucial clause is that for $B_j\phi$:

$$M, w \models B_j\phi \quad \text{iff} \quad \text{for all } u \in W \text{ such that } w \longrightarrow_j u : M, u \models \phi$$

If we assume that the doxastic accessibility relation is serial, transitive, and euclidean, as in the received view, then the logic satisfies the **KD45** axioms (recall Section 2.2.1).

We have explained that beliefs do not remain constant but change in the aftermath of certain actions, inducing model transformations. *Action models* constitute a useful tool to represent complex actions (Baltag et al., 1998). Just as Kripke models capture the uncertainty of agents over possible worlds, action models represent their uncertainty regarding events taking place. This is achieved through accessibility relations imposed on a set of events. These relations may too adhere to algebraic properties. Moreover, each event comes with a *precondition* (a formula in the language) indicating what is required for it to take place. A common example of an action model is that of a *private announcement* whereby only *some* agents are secretly informed whether a coin lies Heads or Tails.

7.1.3. Definition (Action model). An *action model* C is a tuple $\langle E, \{\twoheadrightarrow_j\}_{j\in Ag}, pre\rangle$ where:

- E is a non-empty set of events.
- each $\twoheadrightarrow_j$ is a binary relation on E.
- pre is a precondition function assigning a precondition to each event.

A pair (C, e), consisting of an action model C and an event e in E is called a *pointed action model*. The transformation induced by an action is reflected in *product models*.

7.1.4. Definition (Product model). Let $M = \langle W, \{\longrightarrow_j\}_{j\in Ag}, V\rangle$ be a model and $C = \langle E, \{\twoheadrightarrow_j\}_{j\in Ag}, pre\rangle$ be an action model. The product model $M \otimes C$ is a tuple $\langle W', \{\longrightarrow'_j\}_{j\in Ag}, V'\rangle$ where:

- $W' = \{(w, e) \in W \times E \mid M, w \models pre(e)\}$
- $(w, e) \longrightarrow'_j (w', e')$ iff $(w \longrightarrow_j w'$ and $e \twoheadrightarrow_j e')$
- $V'(w, e) = V(w)$

Dynamic operators of the sort $\langle C, e\rangle$ are introduced to the language of Definition 7.1.2 in order to capture the effect of an event e of an action model C. The semantic interpretation is given with the help of the transformed model.[2]

$$M, w \models \langle C, e\rangle\phi \quad \text{iff} \quad M, w \models pre(e) \text{ and } M \otimes C, (w, e) \models \phi$$

For example, in the cases of private communication, an agent who was informed about the face of the coin (suppose it was Heads up) believes that it lies Heads up, whereas the other agents do not.

In what follows, we will propose our variants of action models which are compatible with impossible-worlds semantics and suitable for representing the various types of reasoning steps.

7.2 The framework

7.2.1 Syntax

The language of our framework is an extension of the simple doxastic language (Definition 7.1.2). Given a set of agents $Ag \neq \varnothing$, its additional components are:

- ▶ Quantitative comparisons between terms that are introduced to compare cognitive costs of rules with the cognitive capacities of agents.
- ▶ Operators A_j, where $j \in Ag$, which indicate the rules that are available to the agent j.

[2] Dual operators, of the form $[C, e]$, can be defined in terms of $\langle C, e\rangle$, as usual in Modal Logic.

- Dynamic operators of the form $\langle \mathsf{C}, e\rangle$, where (C, e) is a pointed action model, designed to capture applications of rules in a multi-agent setting. The details on the design of action models suitable for our purposes will be given in the next section.

To define the language formally we need to introduce certain prerequisite notions: *rules* and (multi-agent) *terms*.

Rules. The dynamic operators $\langle \mathsf{C}, e\rangle$ correspond to actions whose semantic effect is reflected on product updates. In the context of this attempt, we are interested in actions induced by agents taking reasoning steps in a multi-agent environment. It is thus useful to specify what exactly these steps are. We divide them into three main categories, that stand in a clear correspondence to the three topics of interest: deductive reasoning, introspective reasoning, and reasoning about others. Let $\mathcal{L}_\Phi$ be the propositional language based on the set of atoms Φ. Then:

- An *inference rule* ρ is of the form $\{\phi_1, \dots, \phi_n\} \rightsquigarrow \psi$, where $\phi_1, \dots, \phi_n, \psi \in \mathcal{L}_\Phi$. We use $pr(\rho)$ and $con(\rho)$ to abbreviate, respectively, the set of premises and the conclusion of the rule ρ. For example, the following instance of *Disjunctive Syllogism* is such an inference rule: $DS := \{p \vee q, \neg p\} \rightsquigarrow q$.

- An *introspective rule* ρ for agent j is of the form $\{B_j^n \phi\} \rightsquigarrow B_j^{n+1}\phi$, where $n \geq 0$ stands for the number of repetitions of B_j and $\phi \in \mathcal{L}_B$. The premise is taken to be the formula to the left of the rule (the $B_j^n\phi$ assertion), while the conclusion is the formula to the right (the $B_j^{n+1}\phi$ assertion). For example, $\rho := \{B_j\phi\} \rightsquigarrow B_jB_j\phi$ is such an introspective rule.[3]

- An *attributed rule* captures the attribution of an inference rule $\{\phi_1, \dots, \phi_m\} \rightsquigarrow \psi$ to other agents. It is of the form $\{B_{j_1} \dots B_{j_n}\phi_1, \dots, B_{j_1} \dots B_{j_n}\phi_m\} \rightsquigarrow B_{j_1} \dots B_{j_n}\psi$ where $n \geq 1$, $j_1, \dots, j_n \in Ag$ with $j_1 \neq \dots \neq j_n$.[4] The rule that is attributed is called the *object* of the attribution. The left and the right parts are taken to be, respectively, the premises and the conclusion of the attributed rule. They should not be confused with the premises and the conclusion of the object of the attribution. To avoid confusion in dealing with an attributed rule, we sometimes name it by indexing the object by the agents to whom it is attributed. For example, $DS_{j_1} := \{B_{j_1}\neg p, B_{j_1}(p \vee q)\} \rightsquigarrow B_{j_1}q$ is such an attributed rule, where the object DS is attributed to agent j_1.[5]

[3] The introduction of introspective rules is reminiscent of the description of the *Monitoring Mechanism* by Nichols and Stich (2003), which takes a representation ϕ in one's *Belief Box* as input and produces a representation "I believe that ϕ" as output. Notice that what follows can be applied to negative introspective rules too.

[4] Notice that attributed *introspective* rules can be defined similarly.

[5] It is such that if an agent believes the premises of DS_{j_1} (i.e. believes that j_1 believes the premises of DS) then she may come to believe the conclusion (i.e. believe that j_1 believes the conclusion of DS).

We use $\mathcal{L}_R$ to denote the set of all rules. In studying a rule-application in a multi-agent setting, it is also important to specify *who* applies it. We will call the agent who applies a rule the *actor* of the rule. Consider the examples given above. If agent j applies DS, then j is the actor of DS. Similarly, the agent j performing introspection on a belief in ϕ is the actor of that introspective rule. If an agent j applies DS_{j_1}, i.e. attributes DS to j_1, then j is the actor of the attributed rule. We now move to another prerequisite notion, an extension of Definition 3.2.2:

7.2.1. Definition (Multi-agent terms). The set of *multi-agent terms* T_{MA} is defined as $T_{\mathsf{MA}} := \{c_\rho \mid \rho \in \mathcal{L}_R\} \cup \{cp_j \mid j \in Ag\}$. It contains elements for (i) the cognitive costs of rule-applications (of the form c_ρ), and (ii) cognitive capacities of agents (of the form cp_j).

With these in place, we can provide the definition of our language, which is inspired by its single-agent epistemic counterpart appearing in Definition 3.2.3.

7.2.2. Definition (Multi-agent language). The *multi-agent language* $\mathcal{L}_{\mathsf{MA}}$ is an extension of that in Definition 7.1.2, given by:

$$\phi ::= p \mid z_1 s_1 + \ldots + z_n s_n \geq z \mid \neg\phi \mid \phi \wedge \psi \mid A_j \rho \mid B_j \phi \mid \langle \mathsf{C}, e \rangle \phi$$

where $p \in \Phi$, $z_1, \ldots, z_n \in \mathbb{Z}$, $z \in \mathbb{Z}^r$, $s_1, \ldots, s_n \in T_{\mathsf{MA}}$, $\rho \in \mathcal{L}_R$, C is an action model and e an event of it.[6]

Examples of formulas. For instance, $(cp_j \geq c_\rho) \wedge A_j \rho$ is a formula which says that the cognitive capacity of agent j (to which the term cp_j corresponds) is greater or equal than the cognitive cost of a rule ρ (to which the term c_ρ corresponds) and the agent j has ρ available. A formula like $\langle \mathsf{C}, e \rangle B_j \phi$ says that after the event e of the action model C takes place, the agent j believes that ϕ.

7.2.2 Semantics

In this part, we present the building blocks of our semantics. First, we define *resource-sensitive multi-agent models* and provide the semantic interpretations for the static formulas. Second, we design action models for bounded rule-based reasoning. Third, we define suitable product updates between the former and the latter, that will help us interpret the dynamic formulas.

[6] The choice of the number r will be made precise in the next subsection. As in Definition 3.2.3, formulas involving $\leq$, $=$, $-$, $\vee$, $\rightarrow$ can be defined in terms of the rest. For example, a formula of the form $s_1 \geq s_2$ is well-formed: it abbreviates $s_1 + (-1)s_2 \geq \overline{0}$.

A resource-sensitive multi-agent model

We use models that essentially augment multi-agent Kripke models with impossible worlds and cognitive components. They constitute multi-agent doxastic variants of the resource-sensitive models introduced before, having fixed r-many resources Res and the cognitive cost function c, along the lines of Section 3.2.2.

7.2.3. Definition (Resource-sensitive multi-agent model). A *resource-sensitive multi-agent model* (RSMM) is a tuple $\mathsf{M} := \langle W^P, W^I, \{{}^{P}{\longrightarrow}^{P}_{j}\}_{j\in Ag}, \{{}^{P}{\longrightarrow}^{I}_{j}\}_{j\in Ag}, V_P, V_I, R, \{cp_j\}_{j\in Ag}\rangle$ where:

- W^P and W^I are sets of possible and impossible worlds, respectively. Take $W := W^P \cup W^I$.
- each ${}^{P}{\longrightarrow}^{P}_{j}$ is a binary relation on W^P, i.e. ${}^{P}{\longrightarrow}^{P}_{j} \subseteq W^P \times W^P$. Each ${}^{P}{\longrightarrow}^{I}_{j}$ is a binary relation over W^P and W^I, i.e. ${}^{P}{\longrightarrow}^{I}_{j} \subseteq W^P \times W^I$. The doxastic accessibility relation for agent j is given by $\longrightarrow_j := {}^{P}{\longrightarrow}^{P}_{j} \cup {}^{P}{\longrightarrow}^{I}_{j}$.
- $V_P : W^P \to \mathcal{P}(\Phi)$ is a valuation function assigning to each *possible* world, the propositional *atoms* that are true there.
- $V_I : W^I \to \mathcal{P}(\mathcal{L}_{\mathsf{MA}})$ is a valuation function assigning to each *impossible* world, the formulas (*atomic or complex*) that are true there.
- $R : W^P \times Ag \to \mathcal{P}(\mathcal{L}_R)$ is a function that assigns to each pair of a world and an agent the rules that the agent has available there.
- $cp_j \in \mathbb{Z}^r$ stands for the cognitive capacity of agent j, i.e. what j can afford with respect to each resource.[7]

Similarly to Part II, each RSMM comes parameterized by the set of resources Res and the cost function c, but they will not be written down as model components in the interest of simplicity.

Model conditions. To fulfil *Minimal Consistency*, which was motivated in detail in Part II, we ask: $\{\phi, \neg\phi\} \not\subseteq V_I(w)$, for any $w \in W^I$ and $\phi \in \mathcal{L}_{\mathsf{MA}}$. Likewise, to ensure that available inference rules are truth-preserving, we impose *Soundness of inference rules*: for any inference rule ρ, $w \in W^P, j \in Ag$: $\rho \in R(w, j)$ implies $\mathsf{M}, w \models tr(\rho)$ where $tr(\rho) := \bigwedge\limits_{\phi \in pr(\rho)} \phi \to con(\rho)$.[8] Henceforth, whenever we refer to RSMMs, we refer to RSMMs adhering to these conditions.

[7] 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), introspection (Butler, 2013, Chapter 5), and reasoning about others (Apperly et al., 2006; Bradford et al., 2015; Lin et al., 2010).

[8] Beliefs may also be generated from non-truth-preserving inferences, so this restriction depends on the type of reasoning failures one aims to capture. This attempt is mostly oriented at

There are other possible restrictions. For example, in terms of R, we can assume that all introspective rules are available to the agent. It seems plausible that introspection is, in principle, always available. However, an agent might not be able to *afford* taking introspective steps from some point onward, due to their cognitive costs. Therefore, the infinite arrays of positive and negative introspection might fail, but because of the effort they require and not because of an a-priori inability for reflection. In this way, we can reconcile the in principle desirable quality of introspective agents (at least for some theories) and the failure of the axioms of unlimited positive and negative introspection necessitated by descriptive facts (Section 2.3.2). While we do not impose strict restrictions to keep the presentation of the framework as simple and flexible as possible, we put them on the table as reasonable assumptions one might wish to make.[9]

Before we proceed to the dynamics, we define the truth clauses for the static fragment of the language, that is $\mathcal{L}_{\mathsf{MA}}$ without $\langle \mathsf{C}, e\rangle$ operators. To that end, we first need to interpret the terms in T_{MA}. The intuition is that those of the form c_ρ correspond to the cognitive costs of rules and those of the form cp_j correspond to the agent's cognitive capacity. This is why cp_j is used both as a model component and as a term of the language. The use is understood by the context. Thus, given a RSMM M parameterized by Res and c:

7.2.4. Definition (Multi-agent term interpretation). Terms in T_{MA} are interpreted by: $c_\rho^{\mathsf{M}} = c(\rho)$ and $cp_j^{\mathsf{M}} = cp_j$.

Notice that 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. Then, the truth clauses for our static formulas are:

7.2.5. Definition (Multi-agent truth clauses (static fragment)).

$$\begin{array}{llll}
 & \mathsf{M}, w \models p & \text{iff} & p \in V_P(w) \\
 & \mathsf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z & \text{iff} & z_1 s_1^{\mathsf{M}} + \ldots + z_n s_n^{\mathsf{M}} \geq z \\
\text{For } w \in W^P: & \mathsf{M}, w \models \neg\phi & \text{iff} & \mathsf{M}, w \not\models \phi \\
 & \mathsf{M}, w \models \phi \wedge \psi & \text{iff} & \mathsf{M}, w \models \phi \text{ and } \mathsf{M}, w \models \psi \\
 & \mathsf{M}, w \models A_j \rho & \text{iff} & \rho \in R(w, j) \\
 & \mathsf{M}, w \models B_j \phi & \text{iff} & \mathsf{M}, u \models \phi \text{ for all } u \in W \text{ such that } w \longrightarrow_j u \\
\text{For } w \in W^I: & \mathsf{M}, w \models \phi & \text{iff} & \phi \in V_I(w)
\end{array}$$

A formula is *valid in a model* if it is true at all possible worlds, and simply *valid* if it is valid in all models. Notice that the clause for B_j at possible worlds now quantifies over possible *and* impossible worlds, hence leaving room for non-idealized agents.

resource-bounded reasoning and not at logically fallacious inferences, e.g. of the sort emerging in belief bias problems (Evans et al., 1983). These problems, also discussed in Chapter 2, may be better studied in richer frameworks that combine System 1 and System 2 processes, such as that of Chapter 5.

[9] Similar considerations apply to the cost assignment.

Action models

In this framework, we use action models to represent the reasoning steps of agents, falling under any of the three categories of interest, keeping the multi-agent dimension into account. As usual in DEL (Section 7.1), we have to determine the events, the binary relation on events for each agent, and the precondition for each event. The events in the set E can either represent rule-applications (e.g. an application of an instance of Disjunctive Syllogism) or nothing happening. The binary relations $\twoheadrightarrow_j$ and precondition function $pre : E \to \mathcal{L}_{\mathsf{MA}}$ work exactly in the lines of Definition 7.1.3. In addition to these, our action models will have extra components to capture the effect of reasoning steps in RSMMs, since the latter too have additional components compared to plain Kripke models. In detail:

- *Edge-conditions.* Bolander (2018) introduced a generalization of action models, dubbed *edge-conditioned* action models. The rough idea is to attach a "conditionalizing" formula ϕ to each $\twoheadrightarrow_j$ edge of the action model. In a similar spirit, we introduce two types of *edge-condition functions*, $Q_j^P : \twoheadrightarrow_j \to \mathcal{L}_{\mathsf{MA}}$ and $Q_j^I : \twoheadrightarrow_j \to \mathcal{L}_{\mathsf{MA}}$. The former is meant to conditionalize the ${}^P \longrightarrow_j^P$ edges in the product update and the latter the ${}^P \longrightarrow_j^I$ ones. The practical benefit is that we can disallow certain doxastic accessibility edges in the product update, by conditionalizing edges of the action model by $\bot$.

- *Postconditions.* It is useful to introduce postconditions in case an action affects the real world (van Benthem et al., 2006). For our purposes, we need two types of postconditions: (i) a *world postcondition function*: $pos : Ag \times \mathcal{P}(\mathcal{L}_{\mathsf{MA}}) \times E \to \mathcal{P}(\mathcal{L}_{\mathsf{MA}})$ that is introduced to indicate the effect of each event on the valuation of impossible worlds, and (ii) a *capacity postcondition function*, of the form $pos\_cp : Ag \times \mathbb{Z}^r \to \mathbb{Z}^r$, that indicates the effect of actions on cognitive capacities of agents.

- *Label function.* For notational convenience, we use a label function assigning to each event which rule, if any, is applied and who the actors are. For example, if event e_1 stands for an application of ρ only by agent a_1, its label is $(\rho, \{a_1\})$ indicating that the applied rule is ρ and its actor is a_1. If the event represents that nothing happens, its label is $(\varnothing, \varnothing)$: no step occurs and (naturally) no one undertakes it. The label function is of the form $lab : E \to (\mathcal{L}_R \cup \{\varnothing\}) \times \mathcal{P}(Ag)$.

As a result, our action models, dubbed *action models for reasoning*, are defined by:

7.2.6. Definition (Action model for reasoning). An *action model for reasoning* is a tuple $\mathsf{C} := \langle E, \{\twoheadrightarrow_j\}_{j \in Ag}, pre, \{Q_j^P\}_{j \in Ag}, \{Q_j^I\}_{j \in Ag}, pos, pos\_cp, lab \rangle$, where:

- $E, \twoheadrightarrow_j, pre$ are as in Definition 7.1.3.
- $Q_j^P, Q_j^I, pos, pos\_cp, lab$ are as described above.

We now give case studies of action models exemplifying scenarios of deductive inference, introspection, and attribution in a multi-agent context.

Inference by some ($\mathsf{C}_{\mathsf{INF}}$). We explore the case where one agent, who believes the premises of an inference rule ρ, applies the rule, provided that she can. In a multi-agent setting, it is important to also capture what the other agents consider regarding this rule-application. For this case study, the others consider that nothing happens (i.e. the application happens unbeknownst to them). This is captured precisely by the edges ($\twoheadrightarrow_j$) depicted as arrows in Figure 7.1. The set E comprises two events, e_1 to represent the application of ρ by a (hence, $lab(e_1) = (\rho, \{a\})$) and e_0 to represent that nothing happens (hence, $lab(e_0) = (\varnothing, \varnothing)$). The precondition for e_1 is that a believes the premises of the rule, has the rule available and has enough cognitive capacity to apply it. For e_0 it is just $\top$, as nothing happens. The edges are conditionless, i.e. $Q^P_j(e,d) = Q^I_j(e,d) = \top$, for any $j \in Ag$, $e, d \in E$. In other words, the edge-conditions are, in *this* case, trivial and the relations $\twoheadrightarrow_j$ are conditionless, as in plain action models. The postcondition will be used to show that the actor can add the conclusion of ρ in her beliefs, while nothing changes for the other agents. The capacity postcondition is such that only the cognitive capacity of a is reduced (by the cognitive cost of applying the rule ρ), while it remains intact for the rest.

$$pre(e) = \begin{cases} \bigwedge\limits_{\phi \in pr(\rho)} B_a\phi \wedge A_a\rho \wedge (cp_a \geq c_\rho), \text{ if } e = e_1 \\ \top, \text{ if } e = e_0 \end{cases}$$

$$pos(j, X, e) = \begin{cases} X \cup \{con(\rho)\}, \text{if } j = a, e = e_1 \\ X, \text{ otherwise} \end{cases}$$

$$pos\_cp(j, n) = \begin{cases} n - c(\rho), \text{ if } j = a \\ n, \text{ otherwise} \end{cases}$$

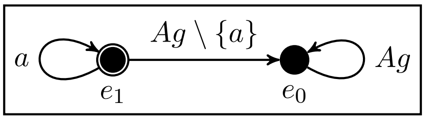


Figure 7.1: The action model $\mathsf{C}_{\mathsf{INF}}$, pointed at e_1, for an application of ρ performed by a while the rest are not aware of it.

Introspection by some ($\mathsf{C}_{\mathsf{INT}}$). This case study is analogous to the previous one: an agent performs an introspective rule unbeknownst to the rest. The action model is as above; only recall that the premise of an introspective rule is a $B^n\phi$ assertion, while the conclusion is a $B^{n+1}\phi$ assertion.

Rule attribution by some ($\mathsf{C}_{\mathsf{ATT}}$). For this case study, one agent (b) reasons about the reasoning of another agent (a), inferring what a would come to believe based on the available information she has. In particular, b attributes inference rule ρ to a, again unbeknownst to the rest. The set E comprises the following events: event e_2 that corresponds to b reasoning about a's reasoning (hence, $lab(e_2) = (\rho_a, \{b\})$), event e_1 that corresponds to agent a applying the rule ρ (hence, $lab(e_1) = (\rho, \{a\})$) and finally, event e_0 that corresponds to the case where nothing happens (hence $lab(e_0) = (\varnothing, \varnothing)$). The precondition for e_2 asks that the agent believes the premises of the attribution, has the inference rule available, and sufficient capacity to attribute it. The values of the action model about e_1 and e_0 are as in the case of a rule-application by a. This is to capture that, in speaking of an *attribution*, apart from b's act of reasoning which refines her own beliefs, b also *thinks* that a engages in an action. The edge-condition functions are non-trivial in this case: namely, $Q^P_b(e_2, e_2) = Q^I_b(e_2, e_1) = \bot$, and $\top$ otherwise.[10] The postconditions are such to show that b enriches her state with the conclusion of the attribution but at the cost of the rule-application.

$$pre(e) = \begin{cases} \bigwedge\limits_{\phi \in pr(\rho)} B_b B_a \phi \wedge A_b \rho \wedge (cp_b \geq c_{\rho_a}), & \text{if } e = e_2 \\ \bigwedge\limits_{\phi \in pr(\rho)} B_a \phi \wedge A_a \rho \wedge (cp_a \geq c_{\rho}), & \text{if } e = e_1 \\ \top, & \text{if } e = e_0 \end{cases}$$

$$pos(j, X, e) = \begin{cases} X \cup \{con(\rho_a)\}, & \text{if } j = b, e = e_2 \\ X \cup \{con(\rho)\}, & \text{if } j = a, e = e_1 \\ X, & \text{otherwise} \end{cases}$$

$$pos\_cp(j, n) = \begin{cases} n - c(\rho_a), & \text{if } j = b \\ n, & \text{otherwise} \end{cases}$$

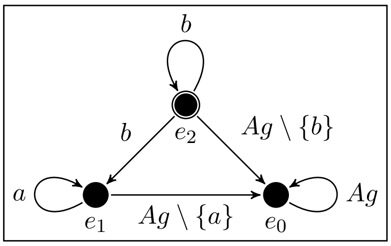


Figure 7.2: The action model $\mathsf{C}_{\mathsf{ATT}}$, pointed at e_2, for an application of ρ_a performed by b while the rest are not aware of it.

More action models can be defined in a similar fashion to capture other scenarios. For instance, we can generalize the constructions above and obtain action models for rule-applications applied by more than one agent (i.e. the set of actors is given by a non-singleton set $B \subseteq Ag$).

Product models

We now define our product models. The RSMMs have additional components compared to Kripke models, like the set of impossible worlds and the cognitive capacity, which are also modified in accordance with the effect of reasoning actions. Roughly, impossible worlds entertained by actors of rules may be eliminated – if their inconsistency is uncovered by a rule-application – or become enriched because, through the application, actors come to believe the conclusion. Moreover, cognitive capacities of agents may be reduced by the appropriate cost. We will describe the components of the product update step by step. First, we need certain abbreviations. Given a RSMM M, $w \in W^P$ and $G \subseteq Ag$:

[10] Disallowing certain edges in the product update serves the following purpose. It will help us capture that the effect of an attribution, as reflected on a product model, is twohold: enriching the *actor's* beliefs but also ensuring that the actor *thinks* her peer (a) enriches *her own* beliefs.

$$\begin{aligned} {}^{P}\!\longrightarrow^{I}_{j}(w) &:= \{u \in W^{I} \mid w\ {}^{P}\!\longrightarrow^{I}_{j} u\} \\ {}^{P}\!\longrightarrow^{I}_{G}(w) &:= \bigcup_{j \in G} {}^{P}\!\longrightarrow^{I}_{j}(w) \\ {}^{P}\!\longrightarrow^{I}_{G} &:= \bigcup_{u \in W^{P}} {}^{P}\!\longrightarrow^{I}_{G}(u) \end{aligned}$$

These abbreviations capture, respectively, which impossible worlds are accessible from w for agent j, for group G, and the ones overall entertained by G. Given a model M and a rule ρ, we also need an abbreviation to talk about impossible worlds that will become inadmissible, given Minimal Consistency, if ρ is applied. That is: $[MC]^{\rho} := \{w \in W^{I} \mid \neg con(\rho) \in V_{I}(w) \text{ or } con(\rho) = \neg\phi, \text{ for some } \phi \in V_{I}(w)\}$. Next, given a model M, an action model C and $e \in E$:

$$[MC]^{e} := \begin{cases} [MC]^{\rho} \cap {}^{P}\!\longrightarrow^{I}_{lab_2(e)}, \text{ if } lab_1(e) = \rho, \text{ for some } \rho \in \mathcal{L}_R \\ \varnothing, \text{ otherwise} \end{cases}$$

This abbreviation allows us to talk about the impossible worlds that will be uncovered as inadmissible by an occurrence of e. For example, if the event represents a ρ-application, then this set of worlds will contain those susceptible to Minimal Consistency that are also entertained by the actors (those who do apply the rule).[11] The components of the product model are designed as follows:

- The new set $(W^{P})'$ consists of pairs of possible worlds and events, such that the world satisfies the precondition of the event (as in Definition 7.1.4).
- The new set $(W^{I})'$ consists of pairs of impossible worlds and events, such that the world is not ruled out by Minimal Consistency. More specifically, if an event e represents a rule-application, the impossible worlds which are paired with it are the ones that survive the rule-application. If an impossible world lies in the doxastic state of an actor who by applying the rule unveils that she initially entertained an inconsistency, then that world will not give rise to such a pair.
- The new relations ${}^{P}\!\longrightarrow^{P'}_{j}$ and ${}^{P}\!\longrightarrow^{I'}_{j}$ are obtained as in Definition 7.1.4, with the extra proviso that the edge-conditions are satisfied.
- The valuation V'_{P} is simply V_{P} restricted to the surviving possible worlds.
- The rationale behind the new valuation V'_{I} is given as follows: if a pair $(w, e) \in (W^{I})'$ lies in the doxastic state of the actor of e, who applies a rule ρ, then its valuation is extended by the conclusion of ρ: such an agent came

[11] For example, if an impossible world w represents $p, p \to q, \neg q$ and is entertained by an agent j, and event e_1 represents the application of $MP = \{p, p \to q\} \rightsquigarrow q$ by j, then w will be contained in $[MC]^{e_1}$. This world will become inadmissible by an e_1 occurrence, because its inconsistency is uncovered by the application of the rule.

to believe the conclusion via the rule-application. Otherwise, the valuation should not be extended, since the states of non-acting agents should not change: *they* do not come to believe the conclusion. This is precisely what the world postcondition function expresses used with the suitable arguments (see Definition 7.2.6).

- ▶ R' is simply R restricted to the surviving worlds.
- ▶ The new cognitive capacity is given through the capacity postcondition. It is such that the capacity of actors is reduced by the cost of the rule-application while the capacity of non-acting agents remains unchanged.

7.2.7. Definition (Product model $\mathsf{M} \otimes \mathsf{C}$). Let M be a RSMM (Definition 7.2.3) and C be an action model for reasoning (Definition 7.2.6). The product model $\mathsf{M} \otimes \mathsf{C}$ is a tuple $\langle (W^P)', (W^I)', \{{}^P\!\!\longrightarrow_j^{P'}\}_{j \in Ag}, \{{}^P\!\!\longrightarrow_j^{I'}\}_{j \in Ag}, V'_P, V'_I, R', \{cp'_j\}_{j \in Ag} \rangle$ where:

- $(W^P)' = \{(w, e) \in W^P \times E \mid \mathsf{M}, w \models pre(e)\}$
- $(W^I)' = \{(w, e) \in W^I \times E \mid w \notin [MC]^e\}$
- Each new doxastic accessibility relation is given by: $\longrightarrow'_j \; := \; {}^P \longrightarrow_j^{P'} \cup {}^P \longrightarrow_j^{I'}$ where:

  ▷ $(w, e) \; {}^P \longrightarrow_j^{P'} (w', e')$ iff $(w \; {}^P \longrightarrow_j^{P} w'$ and $e \twoheadrightarrow_j e'$ and $\mathsf{M}, w \models Q_j^P(e, e'))$

  ▷ $(w, e) \; {}^P \longrightarrow_j^{I'} (w', e')$ iff $(w \; {}^P \longrightarrow_j^{I} w'$ and $e \twoheadrightarrow_j e'$ and $\mathsf{M}, w \models Q_j^I(e, e'))$

- $V'_P(w, e) = V_P(w)$, for $(w, e) \in (W^P)'$
- $V'_I(w, e) = \begin{cases} pos(j, V_I(w), e) \text{ where } j \in lab_2(e), \text{ if } (w, e) \in {}^P \longrightarrow_{lab_2(e)}^{I'} \\ pos(j, V_I(w), e), \text{ where } j \notin lab_2(e), \text{ otherwise} \end{cases}$ for $(w, e) \in (W^I)'$
- $R'((w, e), j) = R(w, j)$, for $(w, e) \in (W^P)'$
- $cp'_j = pos\_cp(j, cp_j)$

The product updates indeed preserve the properties of RSMMs, i.e. if M is a RSMM adhering to our model conditions, and C is an action model for reasoning, defined as above, then the product $\mathsf{M} \otimes \mathsf{C}$ is also a RSMM adhering to our conditions. Product updates allow us to interpret dynamic formulas, i.e. those prefixed by operators of the form $\langle \mathsf{C}, e \rangle$.

7.2.8. Definition (Dynamic truth clauses).

For $w \in W^P$: $\quad \mathsf{M}, w \models \langle \mathsf{C}, e \rangle \phi$ iff $\mathsf{M}, w \models pre(e)$ and $\mathsf{M} \otimes \mathsf{C}, (w, e) \models \phi$

For $w \in W^I$: $\quad \mathsf{M}, w \models \langle \mathsf{C}, e \rangle \phi$ iff $\langle \mathsf{C}, e \rangle \phi \in V_I(w)$

7.2.3 Examples

In this part, we illustrate how the full-fledged semantics works through examples of actions of deductive inference, introspection, and attribution.[12]

7.2.9. Example (**Inference: Mastermind**). We use the *Mastermind* game to illustrate actions of deductive inference. Gierasimczuk et al. (2013) and Zhao et al. (2018) have studied the game to identify which elements are responsible for its cognitive difficulty. The gist of the game is that the *Codemaker* has a secret code of 4 different colours, placed in a particular order, and the *Codebreaker* has to find the secret code. This is achieved by making guesses out of 6 colours, to which the Codemaker provides feedback. Feedback consists in indicating which slots of the guess are correct, in colour alone or in colour *and* position. In this scenario, the Codebreaker (a) gets the feedback depicted in Figure 7.3 from the Codemaker (b).

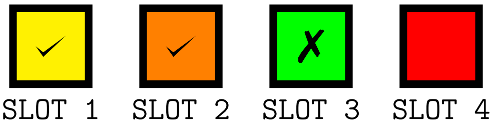


Figure 7.3: The marks in slots 1 and 2 indicate that both the position and the colour of the guess are correct; the mark in slot 3 indicates that neither the position nor the colour is correct; no feedback, as in slot 4, is to say that the colour, but not the position, is correct.

Take r_i, where $i = 1, \dots, 4$, as the propositional atom for the fact "the i-th slot is red" and $r := r_1 \lor r_2 \lor r_3 \lor r_4$ so that r reads "red lies somewhere in the configuration". Fixing *time* and *memory* as our resources of interest, take $cp_a = cp_b = (10, 4)$.[13] The initial situation for agents a and b is given by Figure 7.4. At the beginning, $B_a r$, $B_a \neg r_4$, $B_a \neg(r_1 \lor r_2)$, due to the feedback, but $\neg B_a r_3$, as a has not immediately inferred that the red colour is on the third slot. On the other hand, agent b has access to the correct configuration of colours, therefore $B_b r_3$.

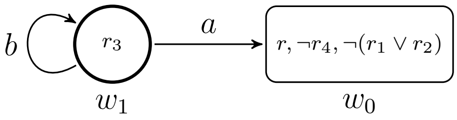


Figure 7.4: The initial RSMM M (pointed at w_1).

Then, agent a unfolds the logical consequences of the feedback, applying successively *Disjunctive Syllogisms* unbeknownst to b; therefore this falls under the $\mathsf{C}_{\mathsf{INF}}$

[12] We follow the conventions of Part II in depicting impossible worlds as rectangles and listing all formulas that are true there. We draw pointed models by depicting the worlds, possible and impossible, that are reachable from the point (and therefore involved in the evaluation of beliefs).

[13] The variant of Mastermind in the learning system *MathGarden* (*rekentuin.nl*), employed by Gierasimczuk et al. (2013) and Zhao et al. (2018), use a timer of "disappearing" coins to monitor an agent's available time against the costs of processing the task.

case. One instance is $DS_1 := \{r, \neg r_4\} \rightsquigarrow r_1 \vee r_2 \vee r_3$ and the other instance is $DS_2 := \{r_1 \vee r_2 \vee r_3, \neg(r_1 \vee r_2)\} \rightsquigarrow r_3$. The agents have them both available, i.e. $DS_1 \in R(w_1, j)$ and $DS_2 \in R(w_1, j)$, for $j = a, b$, and $c(DS_1) = c(DS_2) = (2, 2)$. The action model for the application of DS_1 is C_1 (Figure 7.5) and the action model for the application of DS_2 is C_2 (Figure 7.6).

$$
\begin{aligned}
pre(e) &= \begin{cases} \bigwedge_{\phi \in pr(DS_1)} B_a\phi \wedge A_a(DS_1) \wedge (cp_a \geq c_{DS_1}), & \text{if } e = e_1 \\ \top, \ \text{if } e = e_0 \end{cases} \\
pos(j, X, e) &= \begin{cases} X \cup \{con(DS_1)\}, \text{if } j = a, e = e_1 \\ X, \ \text{otherwise} \end{cases} \\
pos\_cp(j, n) &= \begin{cases} n - c(DS_1), \ \text{if } j = a \\ n, \ \text{otherwise} \end{cases}
\end{aligned}
$$

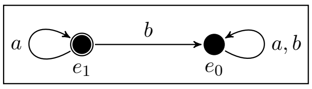


Figure 7.5: The action model C_1, pointed at e_1, for an application of DS_1 performed by a while b is not aware of it.

$$
\begin{aligned}
pre(e) &= \begin{cases} \bigwedge_{\phi \in pr(DS_2)} B_a\phi \wedge A_a(DS_2) \wedge (cp_a \geq c_{DS_2}), & \text{if } e = e_1' \\ \top, \ \text{if } e = e_0' \end{cases} \\
pos(j, X, e) &= \begin{cases} X \cup \{con(DS_2)\}, \text{if } j = a, e = e_1' \\ X, \ \text{otherwise} \end{cases} \\
pos\_cp(j, n) &= \begin{cases} n - c(DS_2), \ \text{if } j = a \\ n, \ \text{otherwise} \end{cases}
\end{aligned}
$$

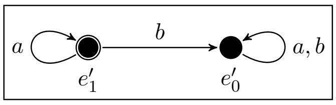


Figure 7.6: The action model C_2, pointed at e_1', for an application of DS_2 performed by a while b is not aware of it.

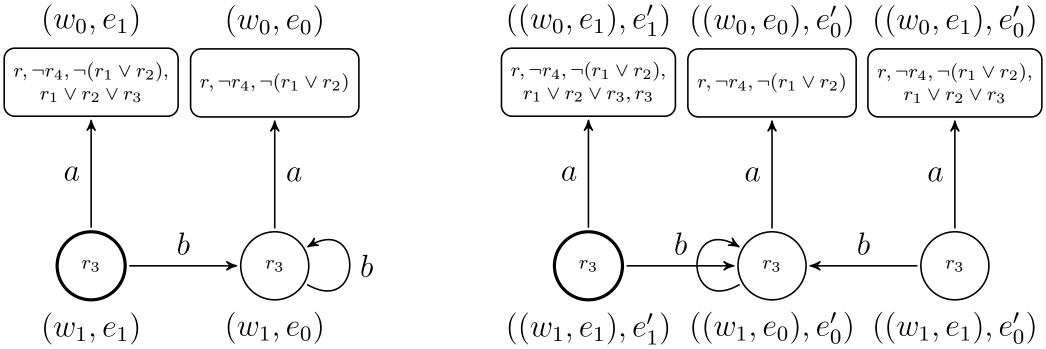

(w_0, e_1)
(w_0, e_0)
$r, \neg r_4, \neg(r_1 \vee r_2), r_1 \vee r_2 \vee r_3$
$r, \neg r_4, \neg(r_1 \vee r_2)$
a
a
r_3
b
r_3
b
(w_1, e_1)
(w_1, e_0)
$((w_0, e_1), e_1')$
$((w_0, e_0), e_0')$
$((w_0, e_1), e_0')$
$r, \neg r_4, \neg(r_1 \vee r_2), r_1 \vee r_2 \vee r_3, r_3$
$r, \neg r_4, \neg(r_1 \vee r_2)$
$r, \neg r_4, \neg(r_1 \vee r_2), r_1 \vee r_2 \vee r_3$
a
a
a
r_3
b
r_3
b
r_3
$((w_1, e_1), e_1')$
$((w_1, e_0), e_0')$
$((w_1, e_1), e_0')$

Figure 7.7: The pointed product models of $\mathsf{M} \otimes \mathsf{C}_1$ (left) and $(\mathsf{M} \otimes \mathsf{C}_1) \otimes \mathsf{C}_2$ (right)

The product model $\mathsf{M} \otimes \mathsf{C}_1$ is built as follows: event e_0 essentially produces a copy of the initial model, thereby capturing the case where nothing new happens. Event e_1 generates a pair for the possible world w_1, because w_1 satisfies its precondition, and a pair for the impossible world w_0, because it is not ruled out by $[MC]^{e_1}$. The doxastic relations are obtained as usual. The valuation for possible worlds and the availability function are unchanged. The valuation for the impossible world (w_0, e_1), which lies in the updated doxastic class of a, is enriched with the conclusion of DS_1, namely $r_1 \vee r_2 \vee r_3$. Notice that b considers possible only the copy of the original model. Moreover, the updated capacities are $cp'_a = (8, 2)$ and $cp'_b = (10, 4)$. The product model $(\mathsf{M} \otimes \mathsf{C}_1) \otimes \mathsf{C}_2$, is obtained analogously. Notice again that the valuation for $((w_0, e_1), e'_1)$, which lies in the doxastic class of a, is enriched with the conclusion of DS_2, namely r_3. Moreover, the final capacities are $cp''_a = (6, 0)$ and $cp''_b = (10, 4)$. As a result, we have:

- $\mathsf{M}, w_1 \models \langle \mathsf{C}_1, e_1\rangle\langle \mathsf{C}_2, e'_1\rangle B_a r_3$ because
  - ▷ $\mathsf{M}, w_1 \models pre(e_1)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}_1, (w_1, e_1) \models pre(e'_1)$
  - ▷ $(\mathsf{M} \otimes \mathsf{C}_1) \otimes \mathsf{C}_2, ((w_1, e_1), e'_1) \models B_a r_3$
- $\mathsf{M}, w_1 \models \langle \mathsf{C}_1, e_1\rangle\langle \mathsf{C}_2, e'_1\rangle \neg B_b B_a r_3$ because
  - ▷ $\mathsf{M}, w_1 \models pre(e_1)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}_1, (w_1, e_1) \models pre(e'_1)$
  - ▷ $(\mathsf{M} \otimes \mathsf{C}_1) \otimes \mathsf{C}_2, ((w_1, e_1), e'_1) \models \neg B_b B_a r_3$

7.2.10. Example (**Introspection: Implicit Bias**). We move to a scenario of an application of an introspective rule, in accordance with the findings discussed in Section 2.3.2. In particular, research on implicit attitudes has unveiled that people holding racist beliefs are often unaware of holding these beliefs, possibly due to the social undesirability these entail (Schwitzgebel, 2010).

For this example, inspired by (Schwitzgebel, 2010, p.532), suppose that agent a holds a belief in p, with p denoting a racist proposition, i.e. $B_a p$. However, a does not believe herself to hold that racist belief, i.e. $\neg B_a B_a p$. The network of a, i.e. agents in $Ag \setminus \{a\}$, are all aware of a's racist belief. Fixing *time* and *memory* in *Res*, take $cp_j = (15, 5)$ for all $j \in Ag$. The initial model is depicted in Figure 7.8. Agent a comes to realize she holds a racist belief through applying the available introspective rule $\rho := \{p\} \rightsquigarrow B_a p$ with $c(\rho) = (10, 3)$. The action model C, an instance of $\mathsf{C}_{\mathsf{INT}}$, is depicted in Figure 7.9.

The product model $\mathsf{M} \otimes \mathsf{C}$, depicted in Figure 7.8 is built as follows: e_0 produces a copy of the initial model to capture the case where nothing new happens. The event e_1 generates a pair for the possible world w_1 because it satisfies its precondition, and a pair for the impossible world w_0 since it is not ruled out by $[MC]^{e_1}$. The valuation for (w_0, e_1), which lies in the doxastic class of a, is enriched

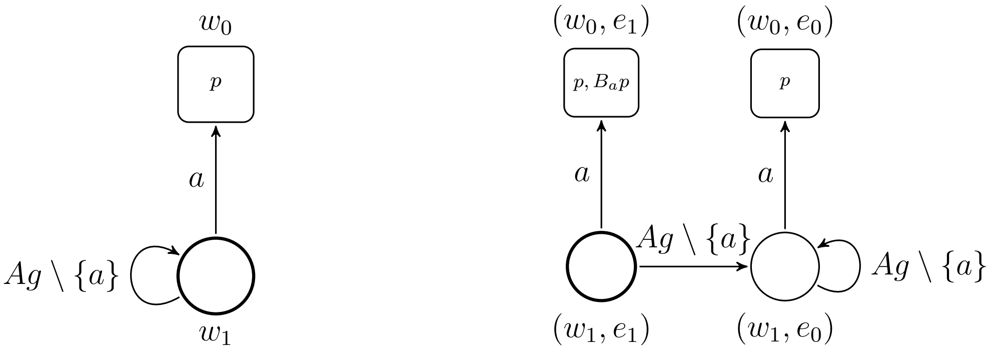


Figure 7.8: The (pointed) initial RSMM M (left) and product model $\mathsf{M} \otimes \mathsf{C}$ (right)

$$pre(e) = \begin{cases} B_a p \wedge A_a \rho \wedge (cp_a \geq c_\rho), \text{ if } e = e_1 \\ \top, \text{ if } e = e_0 \end{cases}$$

$$pos(j, X, e) = \begin{cases} X \cup \{B_a p\}, \text{if } j = a, e = e_1 \\ X, \text{ otherwise} \end{cases}$$

$$pos\_cp(j, n) = \begin{cases} n - c(\rho), \text{ if } j = a \\ n, \text{ otherwise} \end{cases}$$

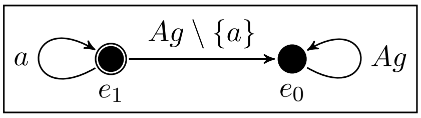


Figure 7.9: The action model C, pointed at e_1, for the introspective rule ρ performed by a while the rest are not aware of it.

with the conclusion of the rule, namely $B_a p$. Agents other than a essentially entertain the copy of the initial model. Moreover, $cp'_a = (5, 2)$ and $cp'_j = (15, 5)$ for $j \neq a$. As a result of the introspective step, a comes to believe she indeed holds a racist belief, but only after some substantial cognitive effort. Others are unaware of this reflection process.

- $\mathsf{M}, w_1 \models \langle \mathsf{C}, e_1 \rangle B_a B_a p$ because
  - ▷ $\mathsf{M}, w_1 \models pre(e_1)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}, (w_1, e_1) \models B_a B_a p$
- $\mathsf{M}, w_1 \models \langle \mathsf{C}, e_1 \rangle \neg B_j B_a B_a p$, for $j \neq a$, because
  - ▷ $\mathsf{M}, w_1 \models pre(e_1)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}, (w_1, e_1) \models \neg B_j B_a B_a p$

7.2.11. Example (**Attribution: Mastermind II**). This example is also based on Mastermind. Besides, the game has been used to test the ability of people to apply ToM (Verbrugge and Mol, 2008). In this scenario, the Codemaker (b), despite observing the actual configuration of colours (the factive information), entertains an impossibility regarding a's beliefs, namely $B_b B_a r$ and $B_b B_a (\neg r_4)$, but $B_b \neg B_a(r_1 \vee r_2 \vee r_3)$. This does justice to the idea that agents are fallible also with regard to higher-order beliefs. Fixing *time* and *memory* in *Res*, take $cp_a = cp_b = (10, 4)$ and $c(DS_{1_a}) = (3, 4)$, where DS_{1_a} amounts to the inference rule DS_1 of Example 7.2.9 attributed to a. The initial model M is depicted in Figure 7.10. However, b can refine her higher-order beliefs provided that she has sufficient resources, by attributing DS_1 to her. The action model C, falling under the $\mathsf{C}_{\mathsf{ATT}}$ type, is depicted in Figure 7.11.

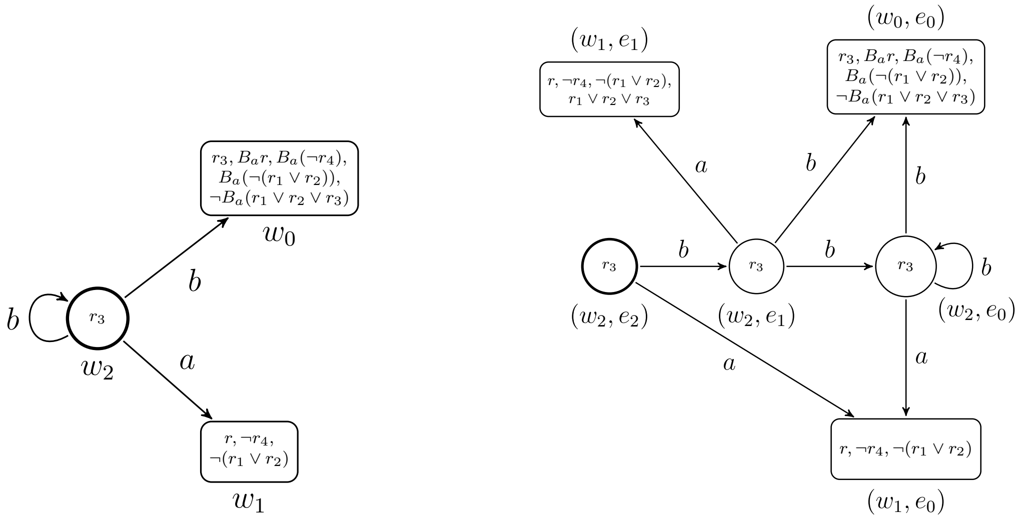

b
r_3
w_2
b
a
$r_3, B_a r, B_a(\neg r_4),$ $B_a(\neg(r_1 \vee r_2)),$ $\neg B_a(r_1 \vee r_2 \vee r_3)$
w_0
$r, \neg r_4,$ $\neg(r_1 \vee r_2)$
w_1
(w_1, e_1)
$r, \neg r_4, \neg(r_1 \vee r_2),$ $r_1 \vee r_2 \vee r_3$
(w_0, e_0)
$r_3, B_a r, B_a(\neg r_4),$ $B_a(\neg(r_1 \vee r_2)),$ $\neg B_a(r_1 \vee r_2 \vee r_3)$
a
b
b
r_3
b
r_3
b
r_3
b
(w_2, e_2)
(w_2, e_1)
(w_2, e_0)
a
a
$r, \neg r_4, \neg(r_1 \vee r_2)$
(w_1, e_0)

Figure 7.10: The initial pointed RSMM (M, w_2) (left) and the pointed product model $(\mathsf{M} \otimes \mathsf{C}, (w_2, e_2))$ (right).

The pointed product model $(\mathsf{M} \otimes \mathsf{C}, (w_2, e_2))$ is also depicted in Figure 7.10. Events e_1 and e_0 generate the part of the new model that corresponds to the outcome of agent a applying DS_1. Event e_2 generates a pair for the possible world w_2, because w_2 satisfies its precondition, and eliminates w_0, because in that case b uncovered the inconsistency in her beliefs. Notice that, from the actual world (w_2, e_2), a considers possible only the copy of the initial model where b has not made the attribution. The capacity of b is reduced by the cost of attributing DS_1 to a ($cp_b = (7, 0)$) while a's remains intact. As a result:

- $\mathsf{M}, w_2 \models \langle \mathsf{C}, e_2 \rangle B_b B_a(r_1 \vee r_2 \vee r_3)$ because
  - ▷ $\mathsf{M}, w_2 \models pre(e_2)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}, (w_2, e_2) \models B_b B_a(r_1 \vee r_2 \vee r_3)$
- $\mathsf{M}, w_2 \models \langle \mathsf{C}, e_2 \rangle \neg B_a B_b B_a(r_1 \vee r_2 \vee r_3)$ because
  - ▷ $\mathsf{M}, w_2 \models pre(e_2)$
  - ▷ $\mathsf{M} \otimes \mathsf{C}, (w_2, e_2) \models \neg B_a B_b B_a(r_1 \vee r_2 \vee r_3)$

$$pre(e) = \begin{cases} \bigwedge_{\phi \in pr(DS_1)} B_b B_a \phi \wedge A_b(DS_1) \wedge (cp_b \geq c_{DS_{1_a}}), \text{ if } e = e_2 \\ \bigwedge_{\phi \in pr(DS_1)} B_a \phi \wedge A_a(DS_1) \wedge (cp_a \geq c_{DS_1}), \text{ if } e = e_1 \\ \top, \text{ for } e = e_0 \end{cases}$$

$$pos(j, X, e) = \begin{cases} X \cup \{con(DS_{1_a})\}, \text{ if } j = b, e = e_2 \\ X \cup \{con(DS_1)\}, \text{ if } j = a, e = e_1 \\ X, \text{ otherwise} \end{cases}$$

$$pos\_cp(j, n) = \begin{cases} n - c(DS_{1_a}), \text{ if } j = b \\ n, \text{ otherwise} \end{cases}$$

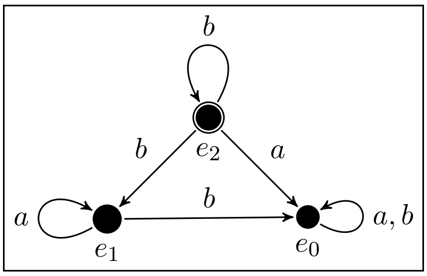


Figure 7.11: The action model C, pointed at e_2, for an application of DS_{1_a} performed by b while a is not aware of it.

7.3 Discussion

In this section, we evaluate the framework, we discuss how the diversity of agents fits under RSMMs, and how this work compares to others in the literature.

Evaluating the framework. Our overarching goal has been to avoid properties of perfect rationality in multi-agent reasoning. Indeed, our agents are not unlimited reasoners who make deductive inferences, perform introspection, or reason about others as if they had endless resources. The design of RSMMs is such that impossible worlds witness the fallibility of agents; an agent might believe ϕ without believing a consequence of it, or without believing that she believes ϕ. Likewise, she might believe that another agent believes ϕ, without thereby attributing a consequence of ϕ to her. However, this does not entail that agents are stuck in an incomplete/inconsistent state and unable to do any better. The agents, provided that they can, take reasoning steps and refine their zero- and higher- order beliefs. The action models allow us to capture these steps and the effort they require. Moreover, we again verify that it is the availability of resources that determines to what extent reasoning evolves; we do not pre-set an arbitrary bound on reasoning but rather we introduce parameters that can embed findings that have long been studied in psychology of reasoning. This serves as a bridge between the formal and the empirical study of reasoning that, despite dealing with the same subject matter, have been developed often in isolation.

The following validities and invalidities back these claims further:

7.3.1. Theorem (Some validities).

1. $\models \bigwedge_{\phi \in pr(\rho)} B_j\phi \wedge A_j\rho \wedge (cp_j \geq c_\rho) \rightarrow \langle \mathsf{C_{INF}}, e_1 \rangle B_j con(\rho)$
   where $lab_1(e_1) = \rho, j \in lab_2(e_1)$

2. $\models B_j B_j^n \phi \wedge A_j\rho \wedge (cp_j \geq c_\rho) \rightarrow \langle \mathsf{C_{INT}}, e_1 \rangle B_j B_j^{n+1}\phi$
   where $lab_1(e_1) = \rho = \{B_j^n\phi\} \rightsquigarrow B_j^{n+1}\phi, j \in lab_2(e_1)$

3. $\models \bigwedge_{\phi \in pr(\rho)} B_b B_a\phi \wedge A_b\rho \wedge (cp_b \geq c_{\rho_a}) \rightarrow \langle \mathsf{C_{ATT}}, e_2 \rangle B_b B_a con(\rho)$
   where $lab_1(e_2) = \rho_a, b \in lab_2(e_2)$

Proof:

We show 1 and 3; the proof for 2 is similar to 1.

1 Take $\mathsf{M}, w \models \bigwedge_{\phi \in pr(\rho)} B_j\phi \wedge A_j\rho \wedge (cp_j \geq c_\rho)$ for arbitrary M and $w \in W^P$. It suffices to show $\mathsf{M}, w \models pre(e_1)$ and $\mathsf{M} \otimes \mathsf{C_{INF}}, (w, e_1) \models B_j con(\rho)$. The former is immediate due to our assumption. For the latter, it suffices to show that $\mathsf{M} \otimes \mathsf{C_{INF}}, (w', e') \models con(\rho)$, for all (w', e') such that $(w, e_1) \longrightarrow_j' (w', e')$. Take arbitrary such (w', e'):

- If $w' \in W^P$, it follows immediately from the assumption, the deductive closure of possible worlds, and the definition of V_P' in Definition 7.2.7.
- If $w' \in W^I$, due to the construction of the new relation in Definition 7.2.7, e' can only be e_1, so $(w, e_1) \ {}^P\!\!\longrightarrow_j^{I'} \ (w', e_1)$ iff $w \ {}^P\!\!\longrightarrow_j^I \ w'$. According to the construction of V_I' in Definition 7.2.7 and the post-condition of $\mathsf{C_{INF}}$, $con(\rho) \in V_I'(w', e')$.

Hence $\mathsf{M} \otimes \mathsf{C_{INF}}, (w', e') \models con(\rho)$, for all (w', e') such that $(w, e_1) \longrightarrow_j' (w', e')$, as desired.

3 Take $\mathsf{M}, w \models \bigwedge_{\phi \in pr(\rho)} B_b B_a\phi \wedge A_b\rho \wedge (cp_b \geq c_{\rho_a})$ for arbitrary M and $w \in W^P$. It suffices to show $\mathsf{M}, w \models pre(e_2)$ and $\mathsf{M} \otimes \mathsf{C_{ATT}}, (w, e_2) \models B_b B_a con(\rho)$. The former is immediate due to our assumption. For the latter, it suffices to show that $\mathsf{M} \otimes \mathsf{C_{ATT}}, (w', e') \models B_a con(\rho)$, for all (w', e') such that $(w, e_2) \longrightarrow_b' (w', e')$. Take arbitrary such (w', e'):

- If $w' \in W^I$: due to the construction of the relation ${}^P\!\!\longrightarrow_j^{I'}$ in Definition 7.2.7 $e' = e_2$. Then, from the definition of the updated V_I' and the postcondition of $\mathsf{C}_{\mathsf{ATT}}$, it follows that $B_a con(\rho) \in V_I'(w', e_2)$.
- If $w' \in W^P$: due to the construction of the relation ${}^P\!\!\longrightarrow_j^{P'}$ in Definition 7.2.7, the only case when $(w, e_2)\ {}^P\!\!\longrightarrow_b^{P'} (w', e')$ is for $w\ {}^P\!\!\longrightarrow_b^P w'$ and $e' = e_1$. We have to show that $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w', e_1) \models B_a con(\rho)$. Take arbitrary (w'', e'') such that $(w', e_1) \longrightarrow_a' (w'', e'')$. There are two cases: either $w'' \in W^P$ or $w'' \in W^I$.
  - ▷ In the latter case, it has to be $w'\ {}^P\!\!\longrightarrow_a^I w''$ and $e'' = e_1$. Then by definition of V_I' in Definition 7.2.7 and the postcondition of $\mathsf{C}_{\mathsf{ATT}}$, it follows that $con(\rho) \in V_I'(w'', e_1)$.
  - ▷ In the former case, we know that $\mathsf{M}, w' \models pre(e_1)$, i.e. $\mathsf{M}, w' \models \bigwedge_{\phi \in pr(\rho)} B_a\phi$, i.e. $\mathsf{M}, w'' \models \bigwedge_{\phi \in pr(\rho)} \phi$. From the deductive closure of possible worlds and the construction of V_P', $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w'', e'') \models con(\rho)$.

  We have established that for all (w'', e'') such that $(w', e_1) \longrightarrow_a' (w'', e'')$, $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w'', e'') \models con(\rho)$, i.e. that $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w', e_1) \models B_a con(\rho)$.

Hence $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w', e') \models B_a con(\rho)$ for all (w', e') such that $(w, e_2) \longrightarrow_b' (w', e')$, i.e. $\mathsf{M} \otimes \mathsf{C}_{\mathsf{ATT}}, (w, e_2) \models B_b B_a con(\rho)$ as desired.

□

The validities show the effect of actions for inference, introspection, and attribution. If an agent believes the premises of a rule, that is cognitively affordable and available to her, then by applying it (and *not automatically*), she comes to believe the conclusion as well. Apart from clearly invalidating the closure properties of logical omniscience and infinite arrays of introspection and attribution, we also present some invalidities that capture the uncertainty of non-actors over the rule-applications of actors. In this sense, they exemplify the multi-agent dimension of this process: as in reality, the interplay of actions has to be continuous for agents to keep track of one another's beliefs.

7.3.2. Theorem (Some invalidities).

1. $\not\models \langle \mathsf{C}_{\mathsf{INF}}, e_1\rangle B_i B_j con(\rho)$ *where* $lab_1(e_1) = \rho, j \in lab_2(e_1), i \notin lab_2(e_1)$
2. $\not\models \langle \mathsf{C}_{\mathsf{INT}}, e_1\rangle B_i B_j B_j^{n+1}\phi$ *where* $lab_1(e_1) = \{B_j^n\phi\} \rightsquigarrow B_j^{n+1}\phi, j \in lab_2(e_1), i \notin lab_2(e_1)$
3. $\not\models \langle \mathsf{C}_{\mathsf{ATT}}, e_2\rangle B_{j_2} B_{j_1} B_{j_2} con(\rho)$ *where* $lab_1(e_2) = \rho_{j_2}, j_1 \in lab_2(e_2), j_2 \notin lab_2(e_2)$

Proof:
We focus on the interesting cases of failure, i.e. those cases where the preconditions hold but the non-prefixed formulas fail in the product models. The counter-examples can be obtained by:

1. $\langle \mathsf{C}_1, e_1 \rangle B_i B_j con(\rho)$ given in Example 7.2.9 for $j = a$, $i = b$ and $\rho = DS_1$ when evaluated at w_1.

2. $\langle \mathsf{C}, e_1 \rangle B_i B_j B_j p$ given in Example 7.2.10 for $j = a$, $i \neq a$ evaluated at w_1.

3. $\langle \mathsf{C}, e_2 \rangle B_{j_2} B_{j_1} B_{j_2} con(\rho)$ given in Example 7.2.11 for $j_1 = b$, $j_2 = a$ and $\rho = DS_1$ evaluated at w_2.

□

Types of agents. Another feature of the framework is the diversity of agents it can encompass. We now discuss in detail different types of agents and on what grounds they fit under this attempt.[14]

Distinctions on grounds of cognitive capacity. The most obvious distinction of agents is in terms of *cp*. For example, an agent might have better memory, that would potentially lead to longer reasoning processes. Drawing on an analogy of Kahneman (1973), Just and Carpenter (1992) have explained that much like a homeowner who draws more units of electrical current than a neighbor, is able to use more units of heating/cooling, one with a larger memory capacity can draw on a larger supply of resources. Evidence of differences in capacity is also documented in studies of individual differences (Just and Carpenter, 1992; Jarrold and Towse, 2006).

Distinctions on grounds of rule-availability. Distinctions can also be made in terms of R. Compare the different performance of experienced and inexperienced reasoners in mathematical reasoning (Weber, 2008; Inglis and Alcock, 2012). Rule-availability may be seen as an *efficiency* indicator, highlighted as another dimension of individual differences, next to capacity. Just and Carpenter (1992) extend the previous analogy: homeowners having more efficient appliances (appliances seen as counterparts of mental processes) manage to produce more units. In our case, an agent with more rules available, e.g. due to experience or training, is in a more advantageous position, compared to someone with an increased capacity but very few tools to make a productive use of it.

Related work. As we saw in Section 2.5.4, some of the challenges we address have already attracted attention, but mostly individually concerning *one* of the aspects we have addressed: deduction, introspection or reasoning about others. Let's see how this attempt compares to these.

[14] For a discussion on diversity of agents in (D)EL, see Liu (2009).

Concerning deductive reasoning, notice that our framework avoids the objections voiced against impossible-worlds semantics on the same grounds as the frameworks of Part II. In other words, we have here generalized the idea behind reasoning-induced updates of resource-sensitive models to further include introspection and attribution – achieved through the design of special action models suitably individuated to each aspect. This does not inhibit, but rather reinforces, the argument for the use of dynamic impossible-worlds semantics against logical omniscience. Furthermore, in contrast to approaches grounded on distinctions of implicit and explicit notions, our approach avoids all forms of logical omniscience and resource-boundedness clearly fits in the picture. We also discussed other dynamic reasoning-oriented approaches. However, instead of a generic notion of reasoning process (be it number of steps, arbitrary rule-applications or time intervals) and the stipulation of an arbitrary cutoff on it, we opted for an elaborate specification of reasoning processes, as envisaged in Chapter 2.

Concerning introspection, while sharing the view of Bonnay and Égré (2009, 2011) and Jago (2009), we specify the role of resources and utilize action models, a key-component of dynamic modelling, to account for the effortful and evolving nature of introspection. Compared to Fervari and Velázquez-Quesada (2019), our model is closer to their *one-step* actions since we too account for *individual* steps – our actions, once taken, do not guarantee full introspection, as we have argued is better in order to import evidence on introspective effort.

Concerning reasoning about others' reasoning, our attempt builds on a different toolbox to that of Alechina et al. (2008), that of modal-doxastic logics (recall Section 2.5.4). This allows us to incorporate the benefits and flexibility of DEL. Moreover, we introduce costs of individual rules as we have argued is necessary for a cognitively plausible framework. Interestingly, the authors describe a problem of *incorrect ascription*. To deal with it, they introduce *reasoning strategies* (well-ordered preferences about how the others reason) and suggest how matching strategies might ensure correct ascriptions. This would be an interesting question to address under the current framework. We also explained how the framework of Balbiani et al. (2019) shares our interest in the dynamics of reasoning. Still, there are differences: our framework does not build on a distinction between background and explicit beliefs, and it does not assume any omniscient notion. The psychologically plausible distinction between long-term and working memory could fit in this setting (see Section 7.5). Also notice that our cognitive parameterization is important in capturing the differential contributions of rules, as rules do not impose equal cognitive burden. Moreover, impossible worlds dispense us from the shortcoming of neighborhood semantics, i.e. closure under equivalence. This is crucial for the study of biases, such as the framing effects. Finally, despite the connections in the spirit of the dynamics, we expect action models to be more flexible in that they can capture a wider variety of informational actions.

7.4 Reduction and axiomatization

This section extends the correspondence between resource-sensitive models and syntactic structures to the multi-agent framework as well. The first part works along the lines presented earlier: we show that the effect of impossible worlds in the interpretation of B can be captured in a (multi-agent) possible-worlds model, provided that suitable auxiliary syntactic functions are introduced. In the second part, we give a sound and complete static axiomatization, through the use of well-known techniques that exploit the correspondence. We move to the third part, the dynamics, trying to provide *reduction axioms*, that, eventually, reduce $\langle \mathsf{C}, e\rangle$-involving formulas to formulas containing no such operator. However, we observe that this common DEL procedure is not straightforward in this case and we explain how this can be overcome, before actually presenting a full axiomatization.

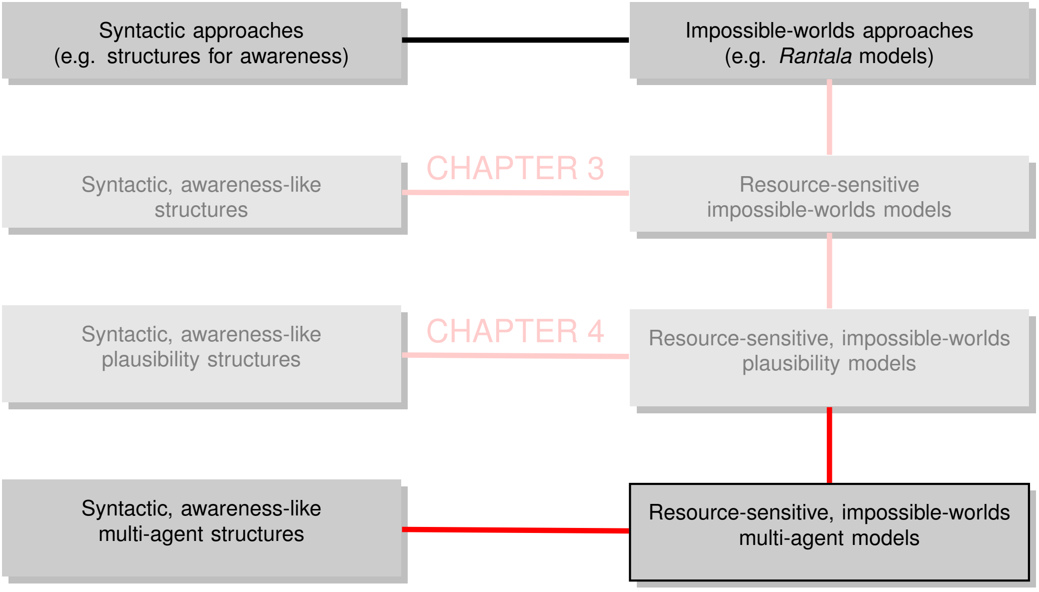

Syntactic approaches
(e.g. structures for awareness)
Impossible-worlds approaches
(e.g. *Rantala* models)
Syntactic, awareness-like
structures
CHAPTER 3
Resource-sensitive
impossible-worlds models
Syntactic, awareness-like
plausibility structures
CHAPTER 4
Resource-sensitive, impossible-worlds
plausibility models
Syntactic, awareness-like
multi-agent structures
Resource-sensitive, impossible-worlds
multi-agent models

Figure 7.12: Extending the reduction to the multi-agent framework

7.4.1 Reduction

Reduced (static) language. We first fix an appropriate language $\mathcal{L}^{red}_{\mathsf{MA}}$ to show that the reduction is successful. Auxiliary operators are introduced to the static fragment of $\mathcal{L}_{\mathsf{MA}}$, in order to discern (syntactically) the effect of possible and impossible worlds in the interpretation of belief. These operators and their semantic interpretations are given below. For $w \in W^P$:

$$\mathsf{M}, w \models L_j\phi \text{ iff } \mathsf{M}, u \models \phi \text{ for all } u \in W^P \text{ such that } w \;{}^{P}\!\!\longrightarrow^{P}_{j}\; u$$
$$\mathsf{M}, w \models I_j\phi \text{ iff } \mathsf{M}, u \models \phi \text{ for all } u \in W^I \text{ such that } w \;{}^{P}\!\!\longrightarrow^{I}_{j}\; u$$

As before, these operators help us break down the B_j operator. We also use $\bot$ as the element of $\mathcal{L}^{red}_{\mathsf{MA}}$, that is never true at any world.

Building the reduced structure. To interpret the I_j auxiliary operators in the reduced structure corresponding to RSMM M, we construct agent-specific *awareness-like functions.*

- $\mathrm{I}_j : W^P \to \mathcal{P}(\mathcal{L}_{\mathsf{MA}})$ such that $\mathrm{I}_j(w) = \bigcap_{u \in {}^P\!\longrightarrow^I_j(w)} V_I(u)$. Intuitively, I_j takes a possible world w and yields the set of those formulas that are true at all impossible worlds in its quantification set (the set of impossible worlds doxastically accessible from w).

7.4.1. Definition (Awareness-like multi-agent structure). Given a RSMM $\mathsf{M} = \langle W^P, W^I, \{{}^P\!\longrightarrow^P_j\}_{j\in Ag}, \{{}^P\!\longrightarrow^I_j\}_{j\in Ag}, V_P, V_I, R, \{cp_j\}_{j\in Ag}\rangle$, its corresponding structure, the *awareness-like multi-agent structure* (ALMS), is $\mathbf{M}_{\mathsf{M}} := \langle \mathrm{W}, \{\longrightarrow^r_j\}_{j\in Ag}, \mathrm{V}, \mathrm{R}, \{cp_j\}_{j\in Ag}, \{\mathrm{I}_j\}_{j\in Ag}\rangle$ where:

$$\begin{array}{cc}
\mathrm{W} = W^P & w \longrightarrow^r_j u \text{ iff } w\ {}^P\!\longrightarrow^P_j u, \text{ for } w, u \in \mathrm{W} \\
\mathrm{V}(w) = V_P(w) \text{ for } w \in \mathrm{W} & \mathrm{R}(w,j) = R(w,j) \text{ for } w \in \mathrm{W}, j \in Ag \\
cp_j \text{ is as in the original} & \mathrm{I}_j \text{ is as explained before}
\end{array}$$

The index M may be omitted if it is easily understood. The clauses based on ALMSs are such that the auxiliary operators I_j are interpreted via the awareness-like functions while L_j behaves as a normal modal operator. By construction of awareness-like functions, Minimal Consistency is inherited by the reduced structure: if a world has impossible worlds accessible from it, then its awareness-like function $\mathrm{I}_j(w)$ will not contain $\{\phi, \neg\phi\}$. Soundness of inference rules is also clearly preserved. The interpretation of terms is as in the RSMMs, since values of capacities and costs are unchanged. The clauses, based on $\mathbf{M}$, are standard for the Boolean connectives. The remaining are given below:

$$\begin{array}{ll}
\mathbf{M}, w \models p \text{ iff } p \in \mathrm{V}(w) & \mathbf{M}, w \models L_j\phi \text{ iff } \mathbf{M}, u \models \phi \text{ for all } u \in \longrightarrow^{\mathrm{r}}_{\mathrm{j}}(w) \\
\mathbf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z \text{ iff } z_1 s_1^{\mathbf{M}} + \ldots + z_n s_n^{\mathbf{M}} \geq z & \mathbf{M}, w \models I_j\phi \text{ iff } \phi \in \mathrm{I}_j(w) \\
\mathbf{M}, w \models A_j\rho \text{ iff } \rho \in \mathrm{R}(w,j) & \mathbf{M}, w \models B_j\phi \text{ iff } \mathbf{M}, w \models L_j\phi \text{ and } \mathbf{M}, w \models I_j\phi
\end{array}$$

We now show that the definition of the ALMSs indeed fulfils its purpose:

7.4.2. Theorem (Reduction). *Given a RSMM* M*, let* $\mathbf{M}$ *be its corresponding ALMS. Then* $\mathbf{M}$ *reduces* M*, i.e. for any* $w \in W^P$ *and formula* $\phi \in \mathcal{L}^{red}_{\mathsf{MA}}$*:* $\mathsf{M}, w \models \phi$ *iff* $\mathbf{M}, w \models \phi$*.*

Proof:
The proof goes by induction on the complexity of ϕ.

- For $\phi := p$: $\mathsf{M}, w \models p$ iff $p \in V_P(w)$ iff $p \in \mathrm{V}(w)$ iff $\mathbf{M}, w \models p$.
- For inequalities, $\neg$, $\wedge$ and A_j, the claim is straightforward because the clauses under $\mathbf{M}$ are given in the same way.
- For $\phi := L_j\psi$: $\mathsf{M}, w \models L_j\psi$ iff $\mathsf{M}, u \models \psi$ for all $u \in W^P$ with $w\ {}^P\!\longrightarrow^P_j\ u$ iff (by I.H.) $\mathbf{M}, u \models \psi$ for all $u \in \mathrm{W}$ such that $w \longrightarrow^r_j\ u$ iff $\mathbf{M}, w \models L_j\psi$.
- For $\phi := I_j\psi$: $\mathsf{M}, w \models I_j\psi$ iff $\mathsf{M}, u \models \psi$ for all $u \in W^I$ such that $w\ {}^P\!\longrightarrow^I_j\ u$ iff $\psi \in V_I(u)$ for all $u \in W^I$ such that $w\ {}^P\!\longrightarrow^I_j\ u$ iff $\psi \in \mathrm{I}_j(w)$ iff $\mathbf{M}, w \models I_j\psi$.
- For $\phi := B_j\psi$: $\mathsf{M}, w \models B_j\psi$ iff $\mathsf{M}, u \models \psi$ for all $u \in W$ such that $w \longrightarrow_j u$. Since $u \in W^P \cup W^I$, this is the case iff $\mathsf{M}, w \models L_j\psi$ and $\mathsf{M}, w \models I_j\psi$. Given the previous steps of the proof, this is the case iff $\mathbf{M}, w \models L_j\psi$ and $\mathbf{M}, w \models I_j\psi$, iff $\mathbf{M}, w \models B_j\psi$.

□

7.4.2 Static axiomatization

Based on the reduction theorem, we provide the static axiomatization:

7.4.3. Definition (Static axiomatization). Λ_{MA} is axiomatized by Table 7.1 and the rules *Modus Ponens*, and *Necessitation*$_{L_j}$ (from ϕ, infer $L_j\phi$).

PC	All instances of classical propositional tautologies
INEQ	All instances of valid formulas about linear inequalities
K_{L_j}	$L_j(\phi \to \psi) \to (L_j\phi \to L_j\psi)$
MC	$I_j\bot \vee (\neg(I_j\phi \wedge I_j\neg\phi))$
SoR	$A_j\rho \to tr(\rho)$, for ρ an inference rule
RED	$B_j\phi \leftrightarrow L_j\phi \wedge I_j\phi$

Table 7.1: The static axioms

INEQ (Fagin et al., 1990; Fagin and Halpern, 1994; Halpern, 2017) is introduced to accommodate the linear inequalities (as in Chapter 3 and Chapter 4). The distribution for L_j is such to show that these operators mimic the behaviour of B_j in the usual multi-agent modal-doxastic logics: these operators quantify over possible worlds only. The axioms for SoR and MC take care of the respective model conditions (Soundness of inference rules and Minimal Consistency). Finally, the last axiom expresses B_j in the terms of the corresponding auxiliary operators.

7.4.4. Theorem (Soundness (static)). Λ_{MA} *is sound with respect to ALMSs.*

Proof:
It suffices to show that our axioms are valid since the rules (*Modus Ponens*, $\textit{Necessitation}_{L_j}$) preserve validity as usual. The claims for PC, INEQ are straightforward, as is for L_j-distribution. The axioms for SoR, MC are valid due to the model conditions. The validity for the last axiom follows from the constructions of the $\mathbf{M}$-semantic clauses for L_j and I_j. □

7.4.5. Theorem (Completeness (static)). Λ_{MA} *is complete with respect to ALMSs.*

Proof:
Because of Theorem 7.4.2, we can employ the technique of canonical model construction (Blackburn et al., 2001). That is, we construct a suitable canonical model and show our truth lemma. Taking (maximal) Λ_{MA}-consistent sets and showing Lindenbaum's lemma go standardly. The canonical model for the logic Λ_{MA} is $\mathcal{M} := \langle \mathcal{W}, \{\longrightarrow^c_j\}_{j \in Ag}, \mathcal{V}, \mathcal{R}, \{cp_j\}_{j \in Ag}, \{\mathcal{I}_j\}_{j \in Ag}\rangle$ where:

$\mathcal{W}$ the set of all maximal Λ_{MA}-consistent sets

$\longrightarrow^c_j$ such that for $w, u \in \mathcal{W}$: $w \longrightarrow^c_j u$ iff $\{\phi \mid L_j\phi \in w\} \subseteq u$

$\mathcal{V}(w) = \{p \mid p \in w\}$, with $w \in \mathcal{W}$

$\mathcal{R}(w, j) = \{\rho \mid A_j\rho \in w\}$, with $w \in \mathcal{W}$

cp_j as in the original

$\mathcal{I}_j(w) = \{\phi \mid I_j\phi \in w\}$, with $w \in \mathcal{W}$

Because of MC and SoR, we ensure that the canonical model has the desired properties: Minimal Consistency and Soundness of inference rules. We then perform induction on the complexity of ψ to show the truth lemma:

$$\mathcal{M}, w \models \psi \text{ iff } \psi \in w$$

- For $\psi := p$: $\mathcal{M}, w \models p$ iff $p \in \mathcal{V}(w)$ iff $p \in w$ by definition of $\mathcal{V}$.
- The claims for Boolean connectives, linear inequalities, and A_j follow directly from the I.H., INEQ, and the construction of the canonical model (properties of maximal consistent sets and $\mathcal{R}$).
- For $\psi := L_j\phi$: $\mathcal{M}, w \models L_j\phi$ iff $\mathcal{M}, u \models \phi$ for all $u \in \mathcal{W}$ such that $w \longrightarrow^c_j u$ iff (by I.H) $\phi \in u$ for all $u \in \mathcal{W}$ such that $w \longrightarrow^c_j u$.
  As a result, we have to show that $L_j\phi \in w$ iff $\phi \in u$ for all $u \in \mathcal{W}$ such that $w \longrightarrow^c_j u$.
  - ▷ The left-to-right direction follows from the definition of $\longrightarrow^c_j$.
  - ▷ The right-to-left direction follows as in Theorem 3.3.5, for L_j essentially acts as the normal modal operator.
- For $\psi := I_j\phi$: $\mathcal{M}, w \models I_j\phi$ iff $\phi \in \mathcal{I}_j(w)$ iff $I_j\phi \in w$ by construction of $\mathcal{I}_j$.
- For $\psi := B_j\phi$: $\mathcal{M}, w \models B_j\phi$ iff $\mathcal{M}, w \models L_j\phi \wedge I_j\phi$ iff $\mathcal{M}, w \models L_j\phi$ and $\mathcal{M}, w \models I_j\phi$ iff (by I.H.) $L_j\phi \in w$ and $I_j\phi \in w$ iff $L_j\phi \wedge I_j\phi \in w$ iff $B_j\phi \in w$.

□

7.4.3 Dynamic axiomatization

Moving to the dynamic part, we look into the behaviour of the reasoning actions under ALMSs. Formulas of the form $\langle \mathsf{C}, e\rangle\phi$ are interpreted as indicated by their original clause (Definition 7.2.8), only now interpreted at the reduced structures corresponding to M and $\mathsf{M} \otimes \mathsf{C}$. More specifically, consider an initial RSMM M and its reduced ALMS $\mathbf{M}$. If a C-update takes place, then we get an updated model $\mathsf{M} \otimes \mathsf{C}$, or for brevity M', and thus an updated ALMS $\mathbf{M}_{\mathsf{M}'}$, or simply $\mathbf{M}'$, corresponding to it. We observe that $\mathbf{M}'$ is such that the updated awareness-like functions I'_j are given in terms of I_j, i.e. the awareness-like functions of $\mathbf{M}$. More specifically, as we show next, the new awareness-like functions are set expressions of the original ones.

The value of an updated awareness-like function is generally given by:

$$\mathrm{I}'_j(w,e) = \bigcap_{(w',e') \in {}^{P}\!\!\longrightarrow^{I'}_{j}(w,e)} V'_I(w',e')$$

Using Definition 7.2.6 and Definition 7.2.7, we compute the values for the updates after the specific case studies of rule-applications discussed in Section 7.2.

- ▶ **After $\mathsf{C}_{\mathsf{INF}}$:** we have two cases; the worlds in the updated structure are either of the form (w, e_1) or (w, e_0), so we have to compute the following:

  ▷ $\mathrm{I}'_j(w,e_1) = \bigcap_{(w',e') \in {}^{P}\!\!\longrightarrow^{I'}_{j}(w,e_1)} V'_I(w',e')$. We have the following cases:

  ∗ If ${}^{P}\!\!\longrightarrow^{I'}_{j}(w,e_1) \neq \varnothing$ and $j \in lab_2(e_1)$, then $(w,e_1)\ {}^{P}\!\!\longrightarrow^{I'}_{j}\ (w',e')$ iff $w\ {}^{P}\!\!\longrightarrow^{I}_{j}\ w'$ and $e' = e_1$. As a result:

$$\begin{aligned}
\mathrm{I}'_j(w,e_1) &= \bigcap_{(w',e_1) \in {}^{P}\!\!\longrightarrow^{I'}_{j}(w,e_1)} V'_I(w',e_1) \\
&= \bigcap_{w' \in {}^{P}\!\!\longrightarrow^{I}_{j}(w)} V'_I(w',e_1) \\
&= \bigcap_{w' \in {}^{P}\!\!\longrightarrow^{I}_{j}(w)} V_I(w') \cup \{con(\rho)\} \\
&= \mathrm{I}_j(w) \cup \{con(\rho)\}
\end{aligned}$$

  ∗ If ${}^{P}\!\!\longrightarrow^{I'}_{j}(w,e_1) \neq \varnothing$ and $j \notin lab_2(e_1)$, then $(w,e_1)\ {}^{P}\!\!\longrightarrow^{I'}_{j}\ (w',e')$ iff $w\ {}^{P}\!\!\longrightarrow^{I}_{j}\ w'$ and $e' = e_0$. As a result:

$$\begin{aligned}
\mathrm{I}'_j(w,e_1) &= \bigcap_{(w',e_0)\in {}^{P}\!\!\longrightarrow_j^{I'}(w,e_1)} V'_I(w',e_0)\\
&= \bigcap_{w'\in {}^{P}\!\!\longrightarrow_j^{I}(w)} V'_I(w',e_0)\\
&= \bigcap_{w'\in {}^{P}\!\!\longrightarrow_j^{I}(w)} V_I(w')\\
&= \mathrm{I}_j(w)
\end{aligned}$$

∗ If ${}^{P}\!\!\longrightarrow_j^{I'}(w,e_1)=\varnothing$, then $\mathrm{I}'_j(w,e_1)=\mathrm{I}_j(w)\cup\overline{\mathrm{I}_j(w)}$

▷ $\mathrm{I}'_j(w,e_0)=\mathrm{I}_j(w)$

- **After** $\mathsf{C}_{\mathsf{INT}}$**:** the updated values for $\mathrm{I}'_j(w,e_1)$ and $\mathrm{I}'_j(w,e_0)$ are as in the previous case.
- **After** $\mathsf{C}_{\mathsf{ATT}}$**:** after attributing ρ to a, we get through the same procedure:

$$\mathrm{I}'_j(w,e_2)=\begin{cases}\mathrm{I}_j(w)\cup\{con(\rho_a)\}, \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_2)\neq\varnothing \text{ and } j\in lab_2(e_2)\\ \mathrm{I}_j(w), \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_2)\neq\varnothing \text{ and } j\notin lab_2(e_2)\\ \mathrm{I}_j(w)\cup\overline{\mathrm{I}_j(w)}, \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_2)=\varnothing\end{cases}$$

$$\mathrm{I}'_j(w,e_1)=\begin{cases}\mathrm{I}_j(w)\cup\{con(\rho)\}, \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_1)\neq\varnothing \text{ and } j\in lab_2(e_1)\\ \mathrm{I}_j(w), \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_1)\neq\varnothing \text{ and } j\notin lab_2(e_1)\\ \mathrm{I}_j(w)\cup\overline{\mathrm{I}_j(w)}, \text{ if } {}^{P}\!\!\longrightarrow_j^{I'}(w,e_1)=\varnothing\end{cases}$$

$$\mathrm{I}'_j(w,e_0)=\mathrm{I}_j(w)$$

We have verified that the updated values of the awareness-like functions are given as set expressions over the original ones. We will use $\mathrm{I}_j^{(\mathsf{C},e)}(w):=\mathrm{I}'_j(w,e)$ to denote the updated values following a (C,e)-update.

In Section 3.3.4 and Section 4.3.3, we explained the practice of providing reduction axioms for dynamic operators in DEL frameworks. In this case, these operators are of the form $\langle\mathsf{C},e\rangle$. However, reducing dynamic formulas involving the auxiliary operator I_j, in terms of which we defined B_j, cannot be straightforwardly obtained in the DEL fashion. This is because the new sets obtained by the update of I_j cannot be described by means of the static language alone. Similar problems are encountered by (Velázquez-Quesada, 2011, Chapter 5), Velázquez-Quesada (2014); in that framework for implicit and explicit beliefs, we find syntactic awareness-like functions, which are expanded after certain actions. The author focuses on action models yielding syntactic functions which are not given arbitrarily, but rather as structured expressions, in turn treatable with a specific static language. We follow a similar procedure, tailored to *our* syntactic functions (I_j). This is because, as shown above, the updated values are too given in terms of the original ones, reflecting the refinement induced by each action. Just to sketch the idea, we extend the static language, essentially re-expressing the auxiliary operators I_j as set-expression operators in the spirit of Velázquez-Quesada (2014) and we then provide reduction axioms that result in a full sound and complete axiomatization.[15]

First, we extend the static language $\mathcal{L}_{\mathsf{MA}}^{red}$ into $\mathcal{L}_{SE}$, in order to generate these expressions; below we give both their semantic interpretations and their corresponding axioms.

7.4.6. Definition (Language $\mathcal{L}_{SE}$). The formulas ϕ and the set expressions over formulas Ω of the language $\mathcal{L}_{SE}$ are given by:

$$\phi ::= p \mid z_1 s_1 + \ldots + z_n s_n \geq z \mid \neg\phi \mid \phi \wedge \psi \mid A_j\rho \mid B_j \mid L_j\phi \mid [\Omega]\phi$$
$$\Omega ::= I_j \mid \{\phi\} \mid \overline{\Omega} \mid \Omega \cup \Omega$$

Notice that the I_j operators are now re-expressed as set-expression operators of the form $[\Omega]$. The added value is that formulas built from set-expression operators allow us to talk not only about the values of awareness-like functions but also about more complex sets built from these values and singleton sets by applying complement and union.

7.4.7. Definition (Truth clauses for new formulas). Given an ALMS $\mathbf{M}$ and $w \in \mathrm{W}$, the truth clauses for the new formulas are given by:

$$\mathbf{M}, w \models [I_j]\phi \text{ iff } \phi \in \mathrm{I}_j(w) \qquad \mathbf{M}, w \models [\{\phi_1\}]\phi_2 \text{ iff } \phi_1 = \phi_2$$
$$\mathbf{M}, w \models [\overline{\Omega}]\phi \text{ iff } \phi \notin \Omega \qquad \mathbf{M}, w \models [\Omega_1 \cup \Omega_2]\phi \text{ iff } \phi \in (\Omega_1 \cup \Omega_2)$$

Note how $[I_j]\phi$ is equivalent to $I_j\phi$.[16] We can further define the intersection and the difference operations as usual. We now present the axioms needed to deal with this extended language, which simply capture the effect of the corresponding set operations, as indicated by Definition 7.4.7.[17]

[15] The fact that updated values are not captured by the static language might look like sufficient reason to abstain from looking into reduction axioms; still, we follow the aforementioned procedure due to the observation that the updates behave in a principled manner. Looking for reduction axioms, in terms of an adapted language, serves as a way to capture this principled behaviour in the logic.

[16] Truth conditions for set expressions under the original model are taken to be the same.

[17] As observed by Velázquez-Quesada (2014), the expressions of complement and union are not strictly necessary as they can be defined in terms of the rest. An extension of the static language with just expressions that allow for expressing syntactic identity between formulas would suffice. Still, their inclusion in the new language will facilitate the reading of the reduction axioms provided afterwards.

$[\{\phi\}]\phi$	$\neg[\{\phi_1\}]\phi_2$, for $\phi_1 \neq \phi_2$
$[\overline{\Omega}]\phi \leftrightarrow \neg[\Omega]\phi$	$[\Omega_1 \cup \Omega_2]\phi \leftrightarrow ([\Omega_1]\phi \vee [\Omega_2]\phi)$

Table 7.2: The axioms for set-expression operators

7.4.8. Theorem (Axiomatization for extended language $\mathcal{L}_{SE}$). *The system given by the axioms of Table 7.2, alongside Definition 7.4.3, is sound and complete for the language $\mathcal{L}_{SE}$ with respect to ALMSs.*

The contribution of the set expressions is towards obtaining reduction axioms for formulas prefixed by $\langle \mathsf{C}, e\rangle$. This is because they allow us to express the effect of action models such as the ones of our case studies. These action models induce set-expression definable updates, in that the updated values of I'_j are set expressions of the original I_j values, as shown before; we denote the set expressions yielding the new values by $[I_j^{(\mathsf{C},e)}]$. To give reduction axioms, we also need to express, in the language, the updated values of capacity and costs:

$$cp_j^{(\mathsf{C},e)} := \begin{cases} cp_j - c_\rho, & \text{when } lab_1(e) = \rho \text{ for some } \rho \in \mathcal{L}_R \text{ and } j \in lab_2(e) \\ cp_j, \text{otherwise} \end{cases} \quad \text{and } c_\rho^{(\mathsf{C},e)} := c_\rho$$

$\langle \mathsf{C}, e\rangle(z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow pre(e) \wedge (z_1 s^{(\mathsf{C},e)} + \ldots + z_n s_n^{(\mathsf{C},e)} \geq z)$	
$\langle \mathsf{C}, e\rangle p \leftrightarrow pre(e) \wedge p$	$\langle \mathsf{C}, e\rangle\neg\phi \leftrightarrow pre(e) \wedge \neg\langle \mathsf{C}, e\rangle\phi$
$\langle \mathsf{C}, e\rangle(\phi \wedge \psi) \leftrightarrow \langle \mathsf{C}, e\rangle\phi \wedge \langle \mathsf{C}, e\rangle\psi$	$\langle \mathsf{C}, e\rangle A_j\rho \leftrightarrow pre(e) \wedge A_j\rho$
$\langle \mathsf{C}, e\rangle\ L_j\phi \leftrightarrow pre(e) \wedge \bigwedge_{e \twoheadrightarrow_j e'}(Q_j^P(e,e') \rightarrow L_j[\mathsf{C}, e']\phi)$	$\langle \mathsf{C}, e\rangle[I_j]\phi \leftrightarrow pre(e) \wedge [I_j^{(\mathsf{C},e)}]\phi$
$\langle \mathsf{C}, e\rangle[\{\phi\}]\phi \leftrightarrow pre(e) \wedge \top$	$\langle \mathsf{C}, e\rangle[\{\phi_1\}]\phi_2 \leftrightarrow pre(e) \wedge \bot$, for $\phi_1 \neq \phi_2$
$\langle \mathsf{C}, e\rangle[\overline{\Omega}]\phi \leftrightarrow \langle \mathsf{C}, e\rangle\neg[\Omega]\phi$	$\langle \mathsf{C}, e\rangle[\Omega_1 \cup \Omega_2]\phi \leftrightarrow \langle \mathsf{C}, e\rangle([\Omega_1]\phi \vee [\Omega_2]\phi)$

Table 7.3: The reduction axioms for dynamic operators

The reduction axioms for the Boolean cases are standard (van Ditmarsch et al., 2007). The axiom for inequalities is such to reflect, with the help of the abbreviations, the resource consumption each action induces. The axiom for rule-availability is not surprising given that it remains unchanged. Since L_j are possible-world quantifying operators that behave normally, the reduction axiom follows the DEL-lines (van Ditmarsch et al., 2007), the difference being the extra requirement of the edge-conditions. The crucial axiom is for I_j-operators. This is given precisely in terms of the set expressions to reflect that the awareness-like functions are updated in a principled way: as specific set expressions of the original ones. After the update, a formula ϕ is in the awareness-like function iff the precondition of e is satisfied and ϕ is in the set expression induced by the update. The axiom for B_j can be derived from the rest while the axioms for the other set expressions are straightforward given their effect already stated in Table 7.2.

7.4.9. Theorem (Full axiomatization). *The axioms and rules of Definition 7.4.3, the axioms of Table 7.2 and Table 7.3, and the rule Action Necessitation (from ϕ, infer $[\mathsf{C}, e]\phi$) provide the full sound and complete system for the language $\mathcal{L}_{SE}$ extended by the dynamic operators with respect to ALMSs and action models for reasoning.*

Proof:
The result follows from Theorem 7.4.4, Theorem 7.4.5 and the validity of the axioms in Table 7.3, which can be easily checked using the truth clauses. □

In short, the procedure has exploited (i) for the static part, the fact that B_j can be broken into two auxiliary operators and the documented correspondence between syntactic and semantic structures against logical omniscience (Wansing, 1990; Fagin et al., 1995), now shown to extend to this framework as well, and (ii) for the dynamics, the fact that the I_j operators can be re-written as set-expression ones that essentially express the same thing, only now allowing us to capture the principled behaviour of our actions.

7.5 Extensions of the framework

In this section, we present some natural extensions of the framework that can be easily realized given the tools hitherto introduced.

Dynamics of interaction. We have seen that next to "internal elucidation", external actions, e.g. public announcements, also enhance an agent's beliefs (van Benthem, 2008c). These fit in our framework, as we explained in Section 4.4. It is precisely the use of novel DEL-like tools that allows for a smooth integration of other established results in this resource-sensitive attempt. For example, our framework can help us study resource-sensitive group notions and it can be combined with plausibility models, as in Chapter 4.

To learn and to forget. The available rules have been assumed to be fixed for each agent and world. However, it can well be that agents *learn* and *forget* rules, hence performing better or worse in reasoning tasks. To that end, we can introduce dynamic operators for learning and forgetting rules, and corresponding model transformations that modify the R function accordingly, as in Section 4.4.3.

Memory. Our treatment of memory as a resource did not discuss distinctions between long-term and working memory (Cowan, 1999), although the reading we assumed throughout the examples is closer to a notion of *working* memory. However, it is interesting to consider both types and impose suitable limitations to each, to ensure that even if notions of background knowledge/beliefs are introduced, they too are resource-sensitive (more specifically with respect to bounds of long-term memory).

Reasoning about others' capabilities. An important element of reasoning about others is reasoning about their capabilities. Alechina et al. (2008) identify this as necessary for correct ascription and Parikh (2007) provides storylines motivating the same point; for example, a sexist agent underestimates the ability of a female agent to derive consequences of a commonly observed fact. We could capture this through B prefixing inequalities: then agents only make attributions if they believe the others can afford it. Still, it is questionable whether agents possess this detailed insight into each other's capacities and into costs of rules. This has been the product of research in psychology of reasoning and serves as a meta-theoretic tool, available to the modeller alone. While we support the need to formally represent agents' reasoning about others' capabilities, we think that this is rarely based on having information about others' specific bounds; rather, the agents have a general attitude about the others' (in)ability to perform certain reasoning steps. This rough "assessment" in multi-agent environments can be better captured through the components of the action models.

7.6 Conclusions

We proposed a uniform way to bridge logical approaches to reasoning and the limitations of humans in deductive and higher-order reasoning. This was achieved by devising RSMMs and action models for reasoning steps of inference, introspection, and attribution, that the agents take to the extent that they can. The combination of impossible-worlds semantics and action models might be of independent interest given the former's use in areas beyond epistemic logic and the latter's popularity in the study of multi-agent dynamics. We also presented connections with syntactic structures whose components, contrary to those of RSMMs, might lack a straightforward reading as features of non-ideal agents but they nonetheless offer the detour towards a sound and complete logic. We have therefore confirmed that this resource-sensitive approach can address challenges beyond logical omniscience, regarding unlimited introspection and reasoning about others, while still preserving the reduction pattern and its contribution. We finally suggested further features this framework can encompass.

This line of work does not constitute a threat to the normative purposes of a logical system, also with respect to multi-agent reasoning: agents still *ought to* derive more information about facts, themselves and others, but only as far as their resources allow. Although multi-agent reasoning involves more than has been discussed (i.e. formation due to observation, memory, and communication in Chapter 6 and rule-based manipulations of higher-order beliefs in Chapter 7) these attempts constitute a step towards the overarching goals of Chapter 2. In the next chapter, we focus on *group reasoning* and, in particular, on whether and how the notion of *distributed knowledge* can be actualized by resource-bounded groups. We will show that the constructions of the current framework are naturally applicable to these questions as well.

Chapter 8

Actualizing distributed knowledge

Epistemic Logic has been used in the study of multi-agent systems, modelling not only the individual knowledge of each agent, but also collective epistemic notions.[1] For example, a group is said to have: *common knowledge* (CK) of ϕ whenever everybody knows that ϕ, everybody knows that everybody knows that ϕ, and so on, *ad infinitum*; and *distributed knowledge* (DK) of ϕ whenever agents can deduce ϕ by pooling their knowledge together. With the tools of DEL, we can further capture the communicative actions giving rise to them, e.g. actions actualizing DK and converting it into CK.

We have seen that EL is often criticized on grounds of idealization: its predictions are practically unattainable by real agents. This has implications for collective notions. It can well be that members of a group do not know all logical consequences of their knowledge (e.g. because of memory overload) or do not take all necessary communicative actions (e.g. because of time pressure). The same constraints apply to higher-order reasoning as agents cannot ascribe knowledge to others at an infinite modal depth. Group reasoning is a dynamic, mixed task that requires actions of both inference and communication. These are not always affordable by human agents, given their cognitive limitations. Therefore, the evolution of group reasoning is also bounded by resources and, even from a normative viewpoint, we should study what can be *feasibly* asked of group members.

This is corroborated by empirical findings (Chapter 2). In deductive reasoning tasks, such as the selection tasks, people often have trouble applying certain inference rules (Section 2.3.1). Other findings study the difficulty of reasoning about others (Section 2.3.3). Group variants of deductive tasks similarly reveal limits in group reasoning (Geil, 1998; Trognon et al., 2011). Nonetheless, they allow us to track which actions underlie successful group performance and the effort they require. It turns out that the distribution of effort among members often yields better performance compared to the individual case.

[1] The chapter is based on Solaki (2020, 2021a).

In light of this, we can revisit group epistemic notions from the perspective of non-ideal agents. Using DEL, we can specify the intertwined effortful actions (communicative and inferential) that refine group knowledge, in accord with empirical facts. Revisiting DK is a first step in this project because of the implicit flavour underlying its understanding as what *would* be known, *if* the agents were to pool their knowledge and deduce information on its basis. In revisiting DK, we need to specify (i) which actions may "actualize" it, i.e. turn it into (explicit) mutual knowledge of the group, and (ii) to what extent these can be undertaken, given that agents are bounded.

The first type of actions is *communicative actions*. Subtleties underpinning the understanding of DK as the outcome of some (unlimited) communication among group members have been discussed by Fagin et al. (1995); Roelofsen (2007); Ågotnes and Wáng (2017). The latter consider the formula $p \wedge \neg K_{a_1} p$: p is true but a_1 does not know it. The formula $D_G(p \wedge \neg K_{a_1} p)$, where G is a group including a_1, is consistent in epistemic logics with D_G operators standing for the DK of G. Yet no communication could render this mutual knowledge of G. The problem lies in that the formula is evaluated in a model that does not explicitly encode the effect of information pooling taking place. The operation introduced by the authors to fill this gap is called *resolution* and it is similar to operations appearing in van Benthem (2011) and Baltag and Smets (2020).

Since our goal is to do justice to non-ideal agents, we should further account for the extent to which resolution can be undertaken. This has implications for the second type of actions too, namely *inferential actions*. There is more than pooling information together that occurs in group deliberations, but unlike communication, the deductive reasoning of group members is usually neglected in multi-agent EL, whereby agents automatically know all consequences of their knowledge. As with communication, we want to encode explicitly the inferential actions of group members, and the extent to which these can be undertaken.

The chapter is organized as follows. In Section 8.1, we present our framework accounting for how agents actualize DK under resource-bounds, using the combination of impossible-worlds semantics and action models proposed in Chapter 7. We illustrate its workings in Section 8.2, and we provide a method for the extraction of a sound and complete axiomatization in Section 8.3.

8.1 The framework

Group reasoning will be approached along the lines of resource-bounded multi-agent reasoning (Chapter 7). We will make use of a special class of RSMMs, suitable for an *epistemic* framework, and of suitable updates due to group communicative/inferential actions. Since our focus is on actualizing DK, our logical language is a variant of $\mathcal{L}_{\mathsf{MA}}$ (Definition 7.2.2) capable of dealing with DK and communicative actions.

8.1.1 Syntax

The logical language of the framework extends that of standard multi-agent epistemic logics. Given a non-empty set of agents Ag, it includes:

- Quantitative comparisons between terms that are introduced to capture cognitive costs of actions (communicative, inferential) with the cognitive capacities of agents.
- Operators D_G, standing for the *distributed knowledge* of group $G \subseteq Ag$.
- Operators A_j, where $j \in Ag$, that indicate the inference rules available to the agent j.
- Operators $\langle \mathsf{R}_G \rangle$, standing for *resolution* of group G, i.e. actions of communication through which the group members pool their knowledge together.
- Operators of the form $\langle \mathsf{C}, e \rangle$, where e is an *event* in *action model* C designed to capture applications of inference rules in a multi-agent setting.

Given the propositional language $\mathcal{L}_\Phi$ based on a set of atoms Φ, we construct the set of inference rules as in Section 7.2.1, which we denote by $\mathcal{L}_R$. The new set of multi-agent terms should be such to represent the cognitive capacities of the agents (as in Definition 7.2.1), but also the cognitive costs of both types of actions (resolution and inference) for *different* types of agents. That is, since we are dealing with resource-bounded *group*-reasoning, a cost will be imposed on the acting agents (as before), but a cost – albeit a minimal one – will also be imposed on the fellow agents who have to wait for their acting peer(s) to undertake the action.

8.1.1. Definition (Group reasoning terms). The set of *group reasoning terms* T_{GR} is defined as $T_{\mathsf{GR}} := \{c_\rho \mid \rho \in \mathcal{L}_R\} \cup \{c_{\rho_1} \mid \rho \in \mathcal{L}_R\} \cup \{c_G \mid G \subseteq Ag\} \cup \{c_{G_1} \mid G \subseteq Ag\} \cup \{cp_j \mid j \in Ag\}$. It contains elements for (i) the cognitive costs of rule-applications for acting agents and non-acting agents (of the form c_ρ and c_{ρ_1}, respectively), (ii) cognitive costs of resolution for acting- and non-acting agents (of the form c_G and c_{G_1}, respectively), (iii) cognitive capacities of agents (of the form cp_j).

We can now formally define our logical language:

8.1.2. Definition (Language). The language $\mathcal{L}_{\mathsf{DK}}$ is given by:

$$\phi ::= p \mid z_1 s_1 + \ldots + z_n s_n \geq z \mid \neg\phi \mid \phi \wedge \psi \mid A_j \rho \mid D_G \phi \mid \langle \mathsf{R}_G \rangle \phi \mid \langle \mathsf{C}, e \rangle \phi$$

where $p \in \Phi$, $z_1, \ldots, z_n \in \mathbb{Z}$, $z \in \mathbb{Z}^r$, $s_1, \ldots, s_n \in T_{\mathsf{GR}}$, $\rho \in \mathcal{L}_R$. The dynamic operators are $\langle \mathsf{R}_G \rangle$ and $\langle \mathsf{C}, e \rangle$, where C is an action model and e an event of it. We will specify the effect of dynamic operators later when presenting the semantics; for now they should be thought as operators for communication and inference respectively.

Examples of formulas. The formula $(cp_j \geq c_\rho) \wedge A_j \rho$ says that (i) the cognitive capacity of agent j (to which the term cp_j corresponds) is greater or equal than the cognitive cost of a rule ρ (to which the term c_ρ corresponds), and (ii) the rule ρ is available to the agent j. Individual knowledge of an agent j is defined in terms of DK as $K_j := D_{\{j\}}$. A formula like $\langle \mathsf{C}, e \rangle K_j \phi$ says that after the event e of the action model C takes place, the agent j knows that ϕ.

8.1.2 Static semantics

We make use of RSMMs (Definition 7.2.3) and suitable model updates, induced by actions of communication (resolution) and inference, corresponding to the effect of our dynamic operators $\langle \mathsf{R}_G \rangle$ and $\langle \mathsf{C}, e \rangle$. We still parameterize the models by a set of r-many resources *Res*, such as *time*, *memory*, etc., and a cognitive cost function c. There is a crucial difference though: the cost function also includes costs of resolution actions. That is, in the context of this attempt, $c : \mathcal{L}_R \cup \mathcal{P}(Ag) \to \mathbb{N}^r$ assigns a *cognitive cost* to (i) each inference rule, (ii) each group, with respect to each resource. That is, cost is a vector (as in Alechina et al. (2009a)), used to indicate the units consumed per resource for actions of inference *and* resolution. We use the notation c_k, $k = 1, \ldots, r$ to refer to the value of the k-th element of the vector and we assume that the first resource, hence the first element of the vector, concerns *time*. This is important because, in speaking of *group* reasoning, we want to account for the time consumed for everyone, acting- or non-acting agents.

Model conditions. We are interested in RSMMs suitable for *epistemic* reasoning. Therefore, we impose *Reflexivity* on the accessibility relation: for every $w \in W^P$, $j \in Ag$: $w \; {}^{P}\!\longrightarrow_j^{P} \; w$, for we accept the factivity of knowledge. In agreement with the arguments provided for the resource-sensitive frameworks of the previous chapters, we again impose *Minimal Consistency*: $\{\phi, \neg\phi\} \not\subseteq V_I(w)$, for any $w \in W^I$ and $\phi \in \mathcal{L}_{\mathsf{DK}}$. Similarly, to ensure that available inference rules are truth-preserving, we impose *Soundness of inference rules*: for $w \in W^P, j \in Ag$: $\rho \in R(w, j)$ implies $\mathsf{M}, w \models tr(\rho)$, where $tr(\rho) := \bigwedge_{\phi \in pr(\rho)} \phi \to con(\rho)$. The RSMMs adhering to these conditions are called *epistemic* RSMMs.

Let's first define the truth clauses for the static fragment, i.e. $\mathcal{L}_{\mathsf{DK}}$ without $\langle \mathsf{R}_G \rangle$ and $\langle \mathsf{C}, e \rangle$ operators. To do that, we first need to interpret the terms in T_{GR}: (a) those of the form c_ρ and c_G correspond to the cognitive costs of rules and group resolution (respectively), (b) those of the form c_{ρ_1} and c_{G_1} correspond to the minimal cognitive costs of rules and group resolution (respectively), affecting the non-acting agents, and (c) those of the form cp_j to the cognitive capacities of agents.

8.1.3. Definition (Interpretation of terms). Given a RSMM M parameterized by *Res* and c, the terms in T_{GR} are interpreted as follows: $c_\rho^{\mathsf{M}} = c(\rho), c_{\rho_1}^{\mathsf{M}} = (c_1(\rho), \ldots, 0)$, $c_G^{\mathsf{M}} = c(G)$, $c_{G_1}^{\mathsf{M}} = (c_1(G), \ldots, 0)$ and $cp_j^{\mathsf{M}} = cp_j$.

We introduce the following abbreviation in order to give the truth clause for DK operators. For $G \subseteq Ag$:

$$\longrightarrow_G \; := \; \bigcap_{j \in G}(\; {}^{P}\!\longrightarrow^{P}_{j} \; \cup \; {}^{P}\!\longrightarrow^{I}_{j} \;)$$

8.1.4. Definition (Static truth clauses).

$$\begin{array}{lllll}
 & \mathsf{M}, w \models p & \text{iff} & p \in V_P(w) \\
 & \mathsf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z & \text{iff} & z_1 s_1^{\mathsf{M}} + \ldots + z_n s_n^{\mathsf{M}} \geq z \\
\text{For } w \in W^P\text{:} & \mathsf{M}, w \models \neg\phi & \text{iff} & \mathsf{M}, w \not\models \phi \\
 & \mathsf{M}, w \models \phi \wedge \psi & \text{iff} & \mathsf{M}, w \models \phi \text{ and } \mathsf{M}, w \models \psi \\
 & \mathsf{M}, w \models A_j \rho & \text{iff} & \rho \in R(w, j) \\
 & \mathsf{M}, w \models D_G \phi & \text{iff} & \mathsf{M}, u \models \phi \text{ for all } u : w \longrightarrow_G u \\
\text{For } w \in W^I: & \mathsf{M}, w \models \phi & \text{iff} & \phi \in V_I(w)
\end{array}$$

Formulas are evaluated directly (i.e. not recursively) by the valuation function at impossible worlds. Notice that the clause for D_G is given through the intersection of relations of G members (as in standard DEL), but it now quantifies over possible *and* impossible worlds, hence leaving room for imperfect agents and groups. A formula is said to be *valid in a model* iff it is true at all possible worlds and *valid in the class of epistemic* RSMMs if it is valid in all epistemic RSMMs.

8.1.3 Resolution

We use resolution as the action that captures how information is pooled by group members, thereby enhancing the group's knowledge. Resolution is understood as publicly known private communication among members (Ågotnes and Wáng, 2017). The resolution of group G induces a model update such that an epistemic relation for a member of G is the intersection of relations of the members of G, and it remains intact for the rest. Moreover, resolution might come at a cost. It can be that "pooling" is effortless, e.g. because information is shared within the group for "free". However, it can also be that adopting a piece of private information through a publicly known action requires effort, e.g. because the group is too big for information to be immediately resolved among all its members or because of the attentional resources this requires, in accord with the distinction of "implicit" and "explicit" informational events (van Benthem, 2008b,c; Velázquez-Quesada, 2009). One way to formally account for this effort is as follows: resolution incurs a non-zero cost to the cognitive capacity for members of G, but also a cost with respect to *time* (and only time) for agents outside G (as time passes while G deliberates). The model update of resolution is below:

8.1.5. Definition (Resolution). Given a RSMM $\mathsf{M} := \langle W^P, W^I, \{ {}^{P}\!\!\longrightarrow^{P}_{j} \}_{j \in Ag}, \{ {}^{P}\!\!\longrightarrow^{I}_{j} \}_{j \in Ag}, V_P, V_I, R, \{cp_j\}_{j \in Ag} \rangle$, the resolution of group G produces the RSMM $\mathsf{M}_G := \langle W^P, W^I, \{ {}^{P}\!\!\longrightarrow^{P'}_{j} \}_{j \in Ag}, \{ {}^{P}\!\!\longrightarrow^{I'}_{j} \}_{j \in Ag}, V_P, V_I, R, \{cp'_j\}_{j \in Ag} \rangle$ where:

$$
{}^{P}\!\longrightarrow_j^{P'} = \begin{cases} \cap_{i\in G}\ {}^{P}\!\longrightarrow_i^{P}, & \text{if } j \in G \\ {}^{P}\!\longrightarrow_j^{P}, & \text{otherwise} \end{cases} \qquad {}^{P}\!\longrightarrow_j^{I'} = \begin{cases} \cap_{i\in G}\ {}^{P}\!\longrightarrow_i^{I}, & \text{if } j \in G \\ {}^{P}\!\longrightarrow_j^{I}, & \text{otherwise} \end{cases}
$$

$$
cp'_j = \begin{cases} cp_j - c(G), & \text{for } j \in G \\ cp_j - (c_1(G), \dots, 0), & \text{otherwise} \end{cases}
$$

The conditions of epistemic RSMMs are preserved by this definition. That is, if M is an epistemic RSMM, then M_G is also an epistemic RSMM. Resolution formulas are then interpreted as follows. For $w \in W$:

$$
\mathsf{M}, w \models \langle \mathsf{R}_G \rangle \phi \text{ iff } \mathsf{M}, w \models (cp_i \geq c_G) \text{ for all } i \in G \text{ and } \mathsf{M}_G, w \models \phi
$$

so the precondition of resolving knowledge within the group G is that the action is cognitively affordable to everyone in the group.[2]

8.1.4 Inference

Action models

We present case studies of action models for reasoning (Definition 7.2.6) capable of representing inferential steps in a group context. Consider, for example, the group selection task. As evinced by the reported dialogues of the participants, Modus Ponens is applied by all agents (Geil, 1998, p.237), (Trognon et al., 2011, pp.15-17). We capture this type of inferential action with the action model below:

Inference by all ($\mathsf{C}_{\mathsf{ALL}}$). This action model captures that *all* agents perform the same reasoning step, the application of a rule ρ, e.g. a Modus Ponens instance. It comprises one reflexive event e_1, and clearly $lab(e_1) = \{\rho, Ag\}$. The edges are condition-less (see $\mathsf{C}_{\mathsf{INF}}$ type of action models for reasoning in Section 7.2.2) therefore the epistemic relations after the action takes place will be produced in no different way than in standard DEL. The precondition is that everybody knows the premises of ρ, has it available, and has enough cognitive capacity to apply it. The postcondition is used to show that agents can add the conclusion in their epistemic state through this rule-application, while the postcondition on capacity reduces it by the cost of ρ.

$$
\begin{aligned}
pre(e_1) &= \bigwedge_{j\in Ag} \bigwedge_{\phi \in pr(\rho)} K_j \phi \wedge \bigwedge_{j \in Ag} A_j \rho \wedge \bigwedge_{j \in Ag} (cp_j \geq c_\rho) \\
pos(j, X, e_1) &= X \cup \{con(\rho)\} \\
pos\_cp(j, n) &= n - c(\rho)
\end{aligned}
$$

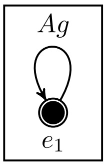


Figure 8.1: The action model for an inference of ρ performed by all.

[2] The interpretation works like those for communication-involving formulas in Section 4.4.

But back to the group selection task: not all agents apply Modus Tollens. In many groups, only one member applies it and figures out that 7 should be turned (Geil, 1998, pp. 238,241). In (Trognon et al., 2011, pp.18-20), some dyads succeed because there is a member with background in logic who has the rule available and affordable, and thus applies it. This is captured by another type of action model:

Inference by some ($\mathsf{C}_{\mathsf{SOME}}$). It is not uncommon that only *some* agents ($G \subset Ag$) perform a rule ρ unbeknownst to agents in $Ag \setminus G$ who do not. The action model comprises two events, e_1 to represent the application of the rule by G (hence, $lab(e_1) = (\rho, G)$) and e_0 to represent that nothing happens (hence, $lab(e_0) = (\varnothing, \varnothing)$). The latter is needed to capture that agents outside of G are uncertain about the content of their peers' action (the rule-application). The precondition for e_1 is that acting agents know the premises of the rule, have the rule available, and have enough cognitive capacity to apply it. For e_0, the precondition is just $\top$, as nothing happens. Again, the edges are condition-less. The postcondition will be used to show that the actors can add the conclusion of ρ in their epistemic state, while nothing changes for the other agents. The cognitive postcondition is such that only the cognitive capacity of the actors is reduced by the cognitive cost of applying ρ, while for the non-actors only time is consumed.

$$pre(e) = \begin{cases} \bigwedge\limits_{a\in G}\ \bigwedge\limits_{\phi\in pr(\rho)} K_a\phi \wedge \bigwedge\limits_{a\in G} A_a\rho \wedge \bigwedge\limits_{a\in G} (cp_a \geq c_\rho), \text{ if } e = e_1 \\ \top, \text{ if } e = e_0 \end{cases}$$

$$pos(j, X, e) = \begin{cases} X \cup \{con(\rho)\}, \text{if } j \in G, e = e_1 \\ X, \text{ otherwise} \end{cases}$$

$$pos\_cp(j, n) = \begin{cases} n - c(\rho), \text{ if } j \in G \\ n - (c_1(\rho), \ldots, 0), \text{ otherwise} \end{cases}$$

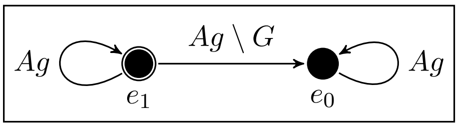


Figure 8.2: The action model (pointed at e_1) for an inference of ρ performed by G unbeknownst to the rest.

Product model. The product between a RSMM M and an action model C is defined by Definition 7.2.7. The operation preserves the properties of epistemic RSMMs. The interpretation for operators $\langle \mathsf{C}, e\rangle$ is given as in Definition 7.2.8:

$$\mathsf{M}, w \models \langle \mathsf{C}, e\rangle\phi \text{ iff } \mathsf{M}, w \models pre(e) \text{ and } \mathsf{M} \otimes \mathsf{C}, (w, e) \models \phi$$

8.2 Discussion

In this section, we see these constructions in action through examples. We also discuss features of the framework, and propose new adaptations of it that naturally fit with its components.

8.2.1 Examples

8.2.1. Example (**Dyad selection task**). For this variant of the selection task (which was initially presented in Section 2.3.1) we focus on two agents, each knowing the visible side of *one* card. The first (a_1) sees the letter card A, and the second (a_2) sees the number card 7.

Language. Denote "card 1 has a vowel" with $\mathtt{v}_1$ and "card 1 has an even number" with $\mathtt{e}_1$. Likewise, $\mathtt{v}_2$ (respectively, $\mathtt{e}_2$) stand for "card 2 has a vowel (even number)". Abbreviate the formulas $\mathtt{v}_i \rightarrow \mathtt{e}_i$ for $i = 1, 2$ with $\mathtt{COND}$. Also, $MP := \{\mathtt{v}_1 \rightarrow \mathtt{e}_1, \mathtt{v}_1\} \rightsquigarrow \mathtt{e}_1$ and $MT := \{\mathtt{v}_2 \rightarrow \mathtt{e}_2, \neg\mathtt{e}_2\} \rightsquigarrow \neg\mathtt{v}_2$.

Initial model. The model representing that a_1 knows the content of the letter card and a_2 knows that of the number card is Figure 8.3 (left). The formulas of $\mathtt{COND}$ are true throughout all worlds. Since agents are fallible, at the beginning they only know what they see (the visible sides) – they have not immediately put their observations together nor have they inferred immediately what lies in the back of the cards. The impossible (incomplete) worlds representing the relevant combinations of letter and number on the first and the second card are:

- w_2: the first card depicts a vowel and the second card an even number.
- w_3: the first card depicts a vowel and the second card an odd number.
- w_4: the first card depicts a consonant and the second card an odd number.

We follow the conventions of Chapter 7 in depicting (pointed) models. We draw impossible worlds as rectangles and write down all formulas true there, to distinguish them from the real (possible) world (w_1), where we write the atoms that are true there, namely $\mathtt{v}_1, \mathtt{e}_1$ (thus $\neg\mathtt{e}_2, \neg\mathtt{v}_2$ are also true as possible worlds are maximal consistent alternatives). The epistemic relations represent the uncertainty of agents with respect to the card they have not seen. Reflexive arrows are omitted for brevity. Moreover, for $Res = \{time, memory\}$, take $cp_{a_1} = (6, 6), cp_{a_2} = (6, 4)$. The rule MP is available to both agents, but MT only to a_1. Finally, $c(MP) = (1, 2)$, $c(MT) = (3, 2)$ as MT is provably more difficult than MP, and $c(G) = (1, 1)$, for the cost of resolution of $G = \{a_1, a_2\}$.

Actions. Afterwards, both agents share their observations. This is captured via resolution. This can be undertaken because $cp_{a_i} \geq c(G)$, for $i = 1, 2$. However, resolution reduces capacities to $(5, 5)$ and $(5, 3)$ respectively. Then, all agents apply MP (captured by an action model of the type $\mathsf{C}_{\mathsf{ALL}}$, called C_1 comprising event e_1), since they both have the rule available and affordable, in accord with the experimental dialogues (Geil, 1998; Trognon et al., 2011). Their capacities become $(4, 3)$ and $(4, 1)$. However, only a_1 applies MT, having the rule available and affordable. This is in accord with the dialogues and it is captured by an action model of the type $\mathsf{C}_{\mathsf{SOME}}$, called C_2 and comprising events e'_1 and e'_0. The actor's capacity finally becomes $(1, 1)$, while a_2's becomes $(1, 1)$ too.

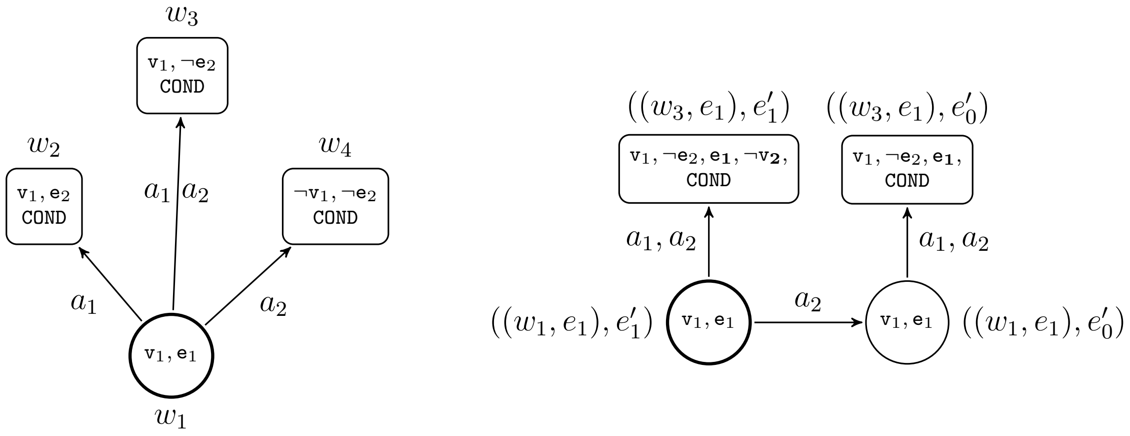

w_3
$\mathtt{v}_1, \neg\mathtt{e}_2$
COND
w_2
$\mathtt{v}_1, \mathtt{e}_2$
COND
a_1
a_2
w_4
$\neg\mathtt{v}_1, \neg\mathtt{e}_2$
COND
a_1
a_2
$\mathtt{v}_1, \mathtt{e}_1$
w_1
$((w_3, e_1), e'_1)$
$\mathtt{v}_1, \neg\mathtt{e}_2, \mathtt{e}_1, \neg\mathtt{v}_2,$
COND
$((w_3, e_1), e'_0)$
$\mathtt{v}_1, \neg\mathtt{e}_2, \mathtt{e}_1,$
COND
a_1, a_2
a_1, a_2
$((w_1, e_1), e'_1)$
$\mathtt{v}_1, \mathtt{e}_1$
a_2
$\mathtt{v}_1, \mathtt{e}_1$
$((w_1, e_1), e'_0)$

Figure 8.3: The initial model M, pointed at w_1, and the updated model M_{fin}, pointed at $((w_1, e_1), e'_1)$.

Final model. The final pointed model is depicted in Figure 8.3 (right). There are reflexive and transitive arrows that are not drawn for the simplicity of the representation. We have $\mathsf{M}_{fin} := (\mathsf{M}_G \otimes \mathsf{C}_1) \otimes \mathsf{C}_2$, resulting from a resolution update (M_G) and then from product updates with C_1 and C_2. It follows that:

$$\mathsf{M}_{fin}, ((w_1, e_1), e'_1) \models K_{a_1}\mathtt{e}_1 \wedge K_{a_1}\neg\mathtt{v}_2 \wedge K_{a_2}\mathtt{e}_1 \wedge \neg K_{a_2}\neg\mathtt{v}_2$$

Therefore:

$$\mathsf{M}, w_1 \models \langle \mathsf{R}_G \rangle \langle \mathsf{C}_1, e_1 \rangle \langle \mathsf{C}_2, e'_1 \rangle (K_{a_1}\mathtt{e}_1 \wedge K_{a_1}\neg\mathtt{v}_2 \wedge K_{a_2}\mathtt{e}_1 \wedge \neg K_{a_2}\neg\mathtt{v}_2)$$

Further development. After another resolution round, a_2 will also come to know $\neg v_2$, since she can afford *that* action (pooling information a_1 derived earlier). This corresponds naturally to the dialogues in (Geil, 1998, pp. 238-240) and (Trognon et al., 2011, pp. 16,19), where the member who figures out that 7 should be turned shares the newly deduced information. Notice that a_2 could use resolution, but not MT; at the end, she did not have to apply MT herself, because her teammate did so, and all she had to do is communicate with her. Had the group not shared their information, they would not have reported the correct solution; had a_2 reasoned alone, her resources would not have allowed her to reach the solution. This illustrates one way in which reasoning in groups facilitates performance in tasks that are more challenging on the individual level.

8.2.2. Example (**Shadow-Box**). This scenario is inspired by the Shadow-Box experiment (Gruber, 1990), which studies the synthesis of disparate points of view by different agents. According to it, there is an object hidden in a box that casts two different shadows on two screens. One subject gets to see shadow 1, and another subject shadow 2. They then have to figure out what the object is. The synthesis can be hard to achieve, depending on the possible configurations of the object, because subjects have the tendency to prioritize their own observations and underestimate each other's input.[3]

Scenario. Suppose that one subject (a_1) only sees the projection on wall 1 (a circle) and the other subject (a_2) only sees the projection on wall 2 (a square). Together they comprise group G.

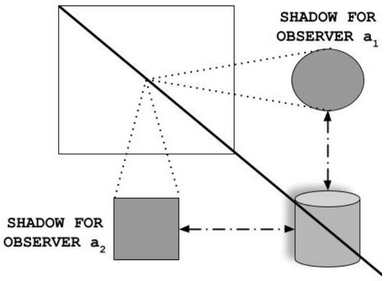


Figure 8.4: The task scenario

Language. Denote "shadow on wall 1 is *c*ircle (*s*quare)" with c_1 (respectively, s_1). Likewise, c_2 (respectively, s_2) stands for "shadow on wall 2 is *c*ircle (*s*quare)". Take *cyli*, *sphe*, *cube* the atoms standing for "the item is a cylinder/sphere/cube". Clearly then: $s_1 \wedge s_2 \to \mathit{cube}$, $c_1 \wedge c_2 \to \mathit{sphe}$, $c_1 \wedge s_2 \to \mathit{cyli}$. Let's refer to these formulas collected together as COND. Take rule Conjunction Introduction $CI := \{c_1, s_2\} \rightsquigarrow c_1 \wedge s_2$ and Modus Ponens $MP := \{c_1 \wedge s_2, c_1 \wedge s_2 \to \mathit{cyli}\} \rightsquigarrow \mathit{cyli}$.

Initial model. Initially the model M is as in Figure 8.5. The real world is w_1, i.e. the hidden object is a cylinder. Agent a_1, who sees a circle shadow, cannot tell apart two cases: the object being a cylinder or a sphere. Agent a_2, who sees a square shadow, cannot tell apart the object being a cylinder or a cube. There are reflexive and transitive arrows, that are omitted for simplicity. Moreover, we fix $Res = \{\mathit{time}, \mathit{memory}\}$ and $c(G) = (1,1)$, $c(CI) = c(MP) = (2,1)$, while $cp_{a_1} = cp_{a_2} = (5,5)$.

Actions. Both agents get together and share their observations. This is taken as a resolution action. Afterwards, they both infer that the item is a cylinder, by using CI (action model C_1, of the type $\mathsf{C}_{\mathsf{ALL}}$, comprising event e_1) and MP (action model C_2 of the type of the type $\mathsf{C}_{\mathsf{ALL}}$, comprising event e'_1). This is possible because they have the rules available and their capacities exceed the costs of the rules.

[3] Addis and Gooding (1999) use computer simulations that see learning as a social process of belief revision between interacting agents, drawing connections between the Shadow-Box and selection tasks. The connection allows for comparisons between the results of Gruber (1990) and applications on a variant of the selection task whereby agents have access to different parts of the world (set of cards), as in Example 8.2.1. This allows us to study reasoning based on disparate observations, which commonly occurs in scientific research and daily life tasks.

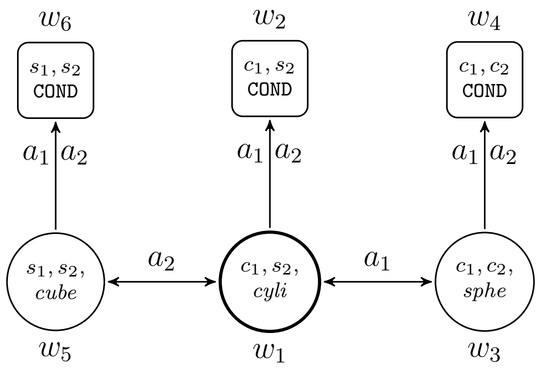


Figure 8.5: The model M after a_1 and a_2 observed the shadows. The impossible worlds represent the agents' observations and prior knowledge on solid objects.

Final model. In Figure 8.6, we give the final (pointed) model $\mathsf{M}_{fin} := (\mathsf{M}_G \otimes \mathsf{C}_1) \otimes \mathsf{C}_2$ after these actions of resolution and inference, in which $cp_{fin}(a_1) = cp_{fin}(a_2) = (0, 2)$. We can verify that $\mathsf{M}_{fin}, ((w_1, e_1), e'_1) \models K_{a_1} cyli \wedge K_{a_2} cyli$, so:

$$\mathsf{M}, w_1 \models \langle \mathsf{R}_G \rangle \langle \mathsf{C}_1, e_1 \rangle \langle \mathsf{C}_2, e'_1 \rangle (K_{a_1} cyli \wedge K_{a_2} cyli)$$

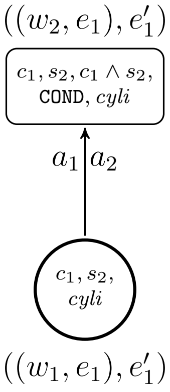


Figure 8.6: The model M_{fin}, pointed at $((w_1, e_1), e'_1)$

8.2.2 Features

We discuss the features of the framework, on basis of the following (in)validities.

8.2.3. Theorem (Some validities in epistemic RSMMs).

1. $\models D_G \langle \mathsf{R}_G \rangle \phi \to \langle \mathsf{R}_G \rangle E_G \phi$ *where* $E_G := \bigwedge_{j \in G} K_j$
2. $\models \bigwedge_{j \in Ag} \bigwedge_{\phi \in pr(\rho)} K_j \phi \wedge \bigwedge_{j \in Ag} A_j \rho \wedge \bigwedge_{j \in Ag} (cp_j \geq c_\rho) \to \langle \mathsf{C}_{\mathsf{ALL}}, e_1 \rangle E_{Ag} con(\rho)$
   where $lab_1(e_1) = \rho$
3. $\models \bigwedge_{\phi \in pr(\rho)} D_G \langle \mathsf{R}_G \rangle \phi \wedge A_j \rho \wedge (cp_j \geq c_G + c_\rho) \to \langle \mathsf{R}_G \rangle \langle \mathsf{C}_{\mathsf{SOME}}, e_1 \rangle K_j con(\rho)$
   where $j \in G, lab_1(e_1) = \rho, j \in lab_2(e_1)$

4. $\models \bigwedge_{\phi \in pr(\rho)} D_{Ag}\langle \mathsf{R}_{Ag}\rangle\phi \wedge \bigwedge_{j \in Ag} A_j\rho \wedge \bigwedge_{j \in Ag}(cp_j \geq c_{Ag} + c_\rho) \to \langle \mathsf{R}_{Ag}\rangle\langle \mathsf{C}_{\mathsf{ALL}}, e_1\rangle E_{Ag}con(\rho)$
where $lab_1(e_1) = \rho$

Proof:

1. Take arbitrary epistemic RSMM M and $w \in W^P$ such that $\mathsf{M}, w \models D_G\langle \mathsf{R}_G\rangle\phi$, i.e. $\mathsf{M}, u \models (cp_j \geq c_G)$ for all $j \in G$ and $\mathsf{M}_G, u \models \phi$, for all $u \in W$ such that $w \longrightarrow_G u$. It suffices to show that $\mathsf{M}_G, w \models E_G\phi$, i.e. that for every $j \in G$, $\mathsf{M}_G, w \models K_j\phi$, i.e. that

   (a) for every $u \in W^P$ such that $w \; {}^{P}\!\longrightarrow_j^{P'} \; u$: $\mathsf{M}_G, u \models \phi$

   (b) for every $u \in W^I$ such that $w \; {}^{P}\!\longrightarrow_j^{I'} \; u$: $\mathsf{M}_G, u \models \phi$

   Because $j \in G$: ${}^{P}\!\longrightarrow_j^{P'} := \bigcap_{j \in G} {}^{P}\!\longrightarrow_j^{P}$ and ${}^{P}\!\longrightarrow_j^{I'} := \bigcap_{j \in G} {}^{P}\!\longrightarrow_j^{I}$. But $\longrightarrow_G := \bigcap_{j \in G}({}^{P}\!\longrightarrow_j^{P} \cup {}^{P}\!\longrightarrow_j^{I})$ and due to our assumption, both desiderata follow.

2. The proof is similar to the ones provided in Theorem 7.3.1.

3. Take arbitrary epistemic RSMM M and $w \in W^P$ such that $\mathsf{M}, w \models \bigwedge_{\phi \in pr(\rho)} D_G \langle \mathsf{R}_G\rangle\phi \wedge A_j\rho \wedge (cp_j \geq c_G + c_\rho)$. Because of item 1 we get that $\mathsf{M}, w \models (cp_i \geq c_G)$ for all $i \in G$ and $\mathsf{M}_G, w \models \bigwedge_{\phi \in pr(\rho)} K_j\phi$. It suffices to show that:

   (a) $\mathsf{M}, w \models (cp_i \geq c_G)$ for all $i \in G$: this follows immediately from the assumption.

   (b) $\mathsf{M}_G, w \models pre(e_1)$, i.e. that $\mathsf{M}_G, w \models \bigwedge_{\phi \in pr(\rho)} K_j\phi \wedge A_j\rho \wedge (cp_j \geq c_\rho)$. We immediately have $\mathsf{M}_G, w \models \bigwedge_{\phi \in pr(\rho)} K_j\phi \wedge A_j\rho$ from the assumption. We then need to ensure that $cp_j^{\mathsf{M}_G} \geq c_\rho^{\mathsf{M}_G}$, i.e. that $cp_j - c(G) \geq c(\rho)$. This also follows from the assumption.

   (c) $\mathsf{M}_G \otimes \mathsf{C}_{\mathsf{SOME}}, (w, e_1) \models K_j con(\rho)$, i.e. that:

      i. all $(w', e') \in (W^P)'$ such that $(w, e_1) \; {}^{P}\!\longrightarrow_j^{P'} \; (w', e')$: $\mathsf{M}_G \otimes \mathsf{C}_{\mathsf{SOME}}, (w', e') \models con(\rho)$. This follows from the assumption, Definition 7.2.7 and the deductive closure of possible worlds.

      ii. all $(w', e') \in (W^I)'$ such that $(w, e_1) \; {}^{P}\!\longrightarrow_j^{I'} \; (w', e')$: $\mathsf{M}_G \otimes \mathsf{C}_{\mathsf{SOME}}, (w', e') \models con(\rho)$. This follows from the assumption, the definition of V_I' in Definition 7.2.7 and the postcondition of $\mathsf{C}_{\mathsf{SOME}}$.

4. The proof is a combination of the previous items (1 and 2).

□

The first validity pertains to the effect of resolution on revisiting notion of DK (in agreement with Ågotnes and Wáng (2017)): after a group resolves their knowledge, ϕ is known by the members. The second captures the effect of actions of inference. The agents do not immediately know all logical consequences of their knowledge: they have to undertake effortful reasoning steps. The other validities encapsulate the interplay of communication and inference: once members resolve their knowledge and come to know the premises, then those who apply the rule, come to know the conclusion as well.

8.2.4. Theorem (Some invalidities in epistemic RSMMs).

1. $\not\models D_G\phi \to \langle \mathsf{R}_G\rangle E_G\phi$

2. $\not\models D_G\langle \mathsf{R}_G\rangle\phi \to \langle \mathsf{R}_G\rangle E_G E_G\phi$

3. $\not\models \bigwedge\limits_{\phi\in pr(\rho)} D_G\phi \to D_G con(\rho)$

4. $\not\models \langle \mathsf{C}_{\mathsf{SOME}}, e_1\rangle K_j con(\rho)$*, where* $\rho = lab_1(e_1)$ *and* $j \notin lab_2(e_1)$

Proof:
The counterexample for 1, given by Ågotnes and Wáng (2017) to motivate the introduction of resolution, also applies here. Counterexamples for 2, 3, and 4 can be easily obtained from Example 8.2.1 and Example 8.2.2. □

The first invalidity unveils the problem behind the traditional understanding of DK (Fagin et al., 1995), also identified by Ågotnes and Wáng (2017). The second invalidity shows that higher orders of knowledge require additional reasoning steps that might not follow from attaining mutual knowledge alone. This departs from literature viewing actualizations of DK as CK, because our attempt focuses on *resource-boundedness*: higher-order knowledge, and hence CK, need extra effort that should not be taken for granted (cf. the higher-order reasoning actions of Chapter 7). The third invalidity shows that DK is not logically closed, therefore actualizing knowledge of logical consequences is not trivial. The fourth invalidity shows that non-acting agents might not come to know logical consequences, even if some of their peers do. This might need yet another round of resolution, exemplifying the continuous and resource-consuming interplay of communication and inference that takes place in reality when non-ideal groups deliberate.

The theorems illustrate how and whether DK is actualized by non-ideal agents. The reasoning actions, their interplay, and the effort they require, allow us to track to which extent a group realizes its potential, instead of pre-setting an arbitrary bound. This reveals the "path" between the implicit notion of DK and the explicit knowledge real groups can achieve.

Recall the case studies: in Example 8.2.2, it was a communicative action (resolution between a_1 and a_2) that provided agents with knowledge of formulas, which in turn functioned as the premises of two rule applications that eventually led them to figure out what the hidden object is. In Example 8.2.1, agent a_1 who initially only knew the content of the first card, was the one to apply Modus Tollens, figured out the other side of the second card, and could subsequently share it with a_2.

Another example of the crucial role of the combination of resolution and inference in actualizing DK is interdisciplinary research. For example, one party might provide input information, but still lack the means (e.g. knowledge of a proof strategy) to make the optimal use of it. Here, this input can become mutual knowledge of the group through resolution, provided researchers come together and make the effort to communicate. Then, the proof strategy (in our terms, the available rules) can build on this information, provided that some agent can apply it, and reach a result that would not have been reached if members worked alone, i.e. their DK would have remained implicit. Scientific quests are largely based on gathering suitable information and deriving further information on its basis. Only when this happens is the scientific potential actualized. However, this process is effortful; it cannot ever be that resolving and deducing comes with no cost in time, memory, attention, etc.

Apart from studying the feasibility of actualizing DK and disentangling communication and inference, the bounds of these processes are also involved in applications that rely on information gathering and processing with cost-accuracy trade-offs. One such example is *recommender systems*, which collect information from multiple and diverse sources to make recommendations of relevant items. The importance of collecting information from various sources is evident, for example, in the development of distributed recommender systems making use of *collaborative filtering* (Bouadjenek et al., 2018). However, next to collecting information, the inferential aspect is also important: as explicit information about the preferences of individuals is generally sparse (often leading to the *cold start problem*), systems often have to resort to implicit information, and perform inferences on basis of them to eventually translate the inputs to meaningful recommendations. However, this process is costly for all parties involved (sending or receiving information) and the accumulation of costs may determine to what extent the process is beneficial and for whom (Vidal, 2004).

8.2.3 Related work

We can break down related work into two aspects, on a par to our dynamics, concerning: (a) the inferential aspect of knowledge, (b) the communicative aspect of actualizing group potential. The former has been discussed in the previous chapter. Moving to aspect (b), Baltag and Smets (2020) and Ågotnes and Wáng (2017) propose actions in accord with the observation of van Benthem (2011): it takes more than communication of formulas expressible in the standard languages to actualize DK. Our resolution action is based on the one of Ågotnes and Wáng (2017) and is similar to a special case of the actions of Baltag and Smets (2020), and to the *communication core* of van Benthem (2011). While this wider variety of actions is compatible with the framework, our dynamics is tailored to bounded agents, explaining how far group reasoning can go. It is precisely this difference in scope that justifies our divergence from studying actualizations of DK as CK. It would also be interesting to connect this resource-sensitive attempt and another generalization of operations for pooling information given by Punčochář and Sedlár (2017): the authors provide an epistemic modality relative to structured communication scenarios as an alternative to distributed knowledge.

The importance of teamwork in realizing goals that would be unattainable if they were approached individually is also witnessed by Dunin-Kęplicz and Verbrugge (2006, 2012). The authors focus on teamwork in multi-agent systems, making use of a multi-modal logical framework that includes both informational and motivational attitudes (such as intentions and commitments). Formalizing collective intentions underlies the construction of different notions of collective commitments, depending on the different kind of agents' awareness necessary to realize a goal under given circumstances. The problem of logical omniscience and the challenges behind reasoning about one's own and others' awareness are acknowledged in the context of these frameworks as well. This is why we find the interplay of *bounded* communicative and inferential actions relevant also in that respect, even if they do not directly touch on the motivational aspect of teamwork.

8.2.4 Variants

There are alternative understandings of the communicative actions, which still are compatible with our framework, e.g. generalizations where agents share all they know with different sets of agents (Baltag and Smets, 2020). This can allow us to break down the effect of resolution into the incremental sharing actions of the members, and study their possibly asymmetrical contribution in actualizing DK.[4]

In particular, Baltag and Smets (2020) introduce a reading map $\alpha : Ag \to \mathcal{P}(Ag)$, mapping each agent $a \in Ag$ to those agents in Ag whose information is accessed by a. A constraint on this map is that $a \in \alpha(a)$, i.e. a's information can always be accessed by herself. The model update after a reading event α is such that the accessibility relation for an agent a is given by the intersection of relations of agents in $\alpha(a)$. There are special cases of such reading actions: cases where one agent's information is accessed by the agents (*tell us all you know*), cases where information is (semi-publicly) shared *between* two groups, cases where information is (semi-publicly) shared *within* one group (like resolution).

Apart from resolution, one can easily include the other reading actions to the

[4] Resolution within a group, and its cost, presuppose a symmetrical *communicative* contribution by the group members (e.g. in Example 8.2.2).

current framework and combine them with group inferential dynamics. The effect on the accessibility relations would be exactly as described above. The difference lies in that our framework is a resource-sensitive one, hence the cognitive capacity might also be modified in the aftermath of a reading action. In the spirit of the effortful resolution update, the reading events should also incur a cost to the agents in the set $\alpha(a)$,[5] and a minimal cost to the remaining agents.

8.3 Reduction and axiomatization

The reduction results of Section 7.4 are also useful to extract a sound and complete resource-bounded logic involving operators for DK. As before, we first show that the effect of impossible worlds in the interpretation of D_G can be captured in a possible-worlds model, provided that suitable syntactic functions are introduced. Second, we obtain a sound and complete static axiomatization. Third, we move to the dynamics. Again, the common DEL procedure of giving reduction axioms, in this case for resolution and inferential actions, is not straightforward, but the issue can be overcome through the technique of Section 7.4.3.

8.3.1 Reduction

Reduced (static) language. To build the reduced static language $\mathcal{L}_{\mathsf{DK}}^{red}$ Take $\longrightarrow_G (w) := \{u \in W \mid w \longrightarrow_G u\}$, which denotes the set the truth clause for D_G quantifies over. Auxiliary operators (L_{D_G}, I_{D_G}) are then introduced to the static fragment of $\mathcal{L}_{\mathsf{DK}}$ to discern syntactically the effect of quantifying over (im)possible worlds in D_G-interpretations. Their interpretations are given below. For $w \in W^P$:

$$\mathsf{M}, w \models L_{D_G}\phi \text{ iff } \mathsf{M}, u \models \phi \text{ for all } u \in W^P \cap \longrightarrow_G (w)$$
$$\mathsf{M}, w \models I_{D_G}\phi \text{ iff } \mathsf{M}, u \models \phi \text{ for all } u \in W^I \cap \longrightarrow_G (w)$$

Building the reduced model. Towards interpreting the auxiliary operators I_{D_G} in a reduced model, we construct suitable *awareness-like functions*:

- $\mathrm{I}_{D_G} : W^P \to \mathcal{P}(\mathcal{L}_{\mathsf{DK}})$ such that $\mathrm{I}_{D_G}(w) = \bigcap_{v \in W^I \cap \longrightarrow_G (w)} V_I(v)$. Intuitively, I_{D_G} takes a possible world and yields the set of formulas true at all impossible worlds in its quantification set (the set of worlds D_G quantifies over).

[5] Assuming that the constraint on reading maps ($a \in \alpha(a)$) is adopted, the information sharing is also costly for a. Notice that the constraint, when adopted by Baltag and Smets (2020), indirectly ensures that the agents have perfect memory. However, in our framework, due to the cost function, we would avoid this feature even when adopting the constraint, as should arguably be for non-ideal agents.

8.3.1. Definition (Epistemic awareness-like multi-agent structure). Given an epistemic RSMM $\mathsf{M} = \langle W^P, W^I, \{^P\!\!\longrightarrow^P_j\}_{j\in Ag}, \{^P\!\!\longrightarrow^I_j\}_{j\in Ag}, V_P, V_I, R, \{cp_j\}_{j\in Ag}\rangle$, its epistemic *awareness-like multi-agent structure* (epistemic ALMS) is $\mathbf{M}_{\mathsf{M}} := \langle \mathrm{W}, \{\longrightarrow^r_j\}_{j\in Ag}, \mathrm{V}, \mathrm{R}, \{cp_j\}_{j\in Ag}, \mathrm{I}_{D_G}\rangle$ with:

$$\begin{array}{cc}
\mathrm{W} = W^P & \mathrm{R}(w,j) = R(w,j) \text{ for } w \in \mathrm{W} \\
w\longrightarrow^r_j u \text{ iff } w\ ^P\!\!\longrightarrow^P_j\ u, \text{ for } w, u \in \mathrm{W} & cp_j \text{ is as in the original} \\
\mathrm{V}(w) = V_P(w) \text{ for } w \in \mathrm{W} & \mathrm{I}_{D_G} \text{ as explained before}
\end{array}$$

The index M may be omitted if it is easily understood. The clauses based on the reduced model are such that the I_{D_G}-operators are interpreted via the awareness-like functions. Again, Minimal Consistency is inherited by the reduced model: for no $w \in \mathrm{W}$, $G \subseteq Ag$, is it the case that $\{\phi, \neg\phi\} \subseteq \mathrm{I}_{D_G}(w)$, when w has impossible worlds accessible from it. Soundness of inference rules and Reflexivity are also clearly preserved. Moreover, take $\longrightarrow^r_j(w) := \{u \in \mathrm{W} \mid w\longrightarrow^r_j u\}$ now based on the new relation $\longrightarrow^r_j$. The interpretation of terms is as in the original, since the values of capacities and costs are unchanged. The semantic clauses, based on $\mathbf{M}$, are standard for the Boolean connectives. The remaining:

$$\begin{array}{ll}
\mathbf{M}, w \models p \text{ iff } p \in \mathrm{V}(w) & \mathbf{M}, w \models L_{D_G}\phi \text{ iff } \mathbf{M}, u \models \phi \text{ for all } u \in \bigcap_{j\in G} \longrightarrow^{\mathrm{r}}_{\mathrm{j}}(w) \\
\mathbf{M}, w \models z_1 s_1 + \ldots + z_n s_n \geq z \text{ iff } z_1 s_1^{\mathbf{M}} + \ldots + z_n s_n^{\mathbf{M}} \geq z & \mathbf{M}, w \models I_{D_G}\phi \text{ iff } \phi \in \mathrm{I}_{D_G}(w) \\
\mathbf{M}, w \models A_j\rho \text{ iff } \rho \in \mathrm{R}(w,j) & \mathbf{M}, w \models D_G\phi \text{ iff } \mathbf{M}, w \models L_{D_G}\phi \text{ and } \mathbf{M}, w \models I_{D_G}\phi
\end{array}$$

We now show that the definition of the epistemic ALMSs indeed fulfils its purpose:

8.3.2. Theorem (Reduction). *Given an epistemic RSMM* M, *let* $\mathbf{M}$ *be its corresponding epistemic ALMS. Then* $\mathbf{M}$ *is a reduction of* M, *i.e. for any* $w \in W^P$ *and formula* $\phi \in \mathcal{L}^{red}_{\mathsf{DK}}$: $\mathsf{M}, w \models \phi$ *iff* $\mathbf{M}, w \models \phi$.

Proof:
The proof goes by induction on the complexity of ϕ. □

8.3.2 Static axiomatization

Based on the reduction theorem, we provide the static axiomatization:

8.3.3. Definition (Static axiomatization). Λ_{DK} is axiomatized by Table 8.1 and the rules *Modus Ponens*, *Necessitation*$_{L_{D_G}}$ (from ϕ, infer $L_{D_G}\phi$).

The basic axioms PC and INEQ are as in Section 7.4.2. The axioms for L_{D_G} ($\mathsf{K}_{D_G}, \mathsf{T}_{D_G}, \mathsf{MON}$) mimic the behaviour of D_G-involving axioms in the standard logics with DK (Fagin et al., 1992, 1995; Gerbrandy, 1999). The axioms MC and SoR take care of the model conditions: Minimal Consistency and Soundness of inference rules, respectively. Finally, the axiom RED reduces D_G in terms of the corresponding auxiliary operators.

PC	All instances of classical propositional tautologies
INEQ	All instances of valid formulas about linear inequalities
K_{D_G}	$L_{D_G}(\phi \to \psi) \to (L_{D_G}\phi \to L_{D_G}\psi)$
T_{D_G}	$L_{D_G}\phi \to \phi$
MON	$L_{D_G}\phi \to L_{D_H}\phi$, if $G \subseteq H$
	$I_{D_G}\phi \to I_{D_H}\phi$, if $G \subseteq H$
MC	$I_{D_G}\bot \vee (\neg(I_{D_G}\phi \wedge I_{D_G}\neg\phi))$
SoR	$A_j\rho \to tr(\rho)$
RED	$D_G\phi \leftrightarrow L_{D_G}\phi \wedge I_{D_G}\phi$

Table 8.1: The static axioms

8.3.4. Theorem (Λ_{DK} soundness and completeness). Λ_{DK} *is sound and complete with respect to epistemic ALMSs.*

Proof:

- Soundness: it suffices to show that our axioms are valid since the rules (*Modus Ponens*, *Necessitation*$_{L_{D_G}}$) preserve validity as usual. The claims for PC, INEQ are straightforward, T_{D_G} is valid given the reflexivity of the accessibility relations. The axioms for MC and SoR are valid due to Minimal Consistency and Soundness of rules. It is easy to check that the axioms for monotonicity (MON) are valid due to the interpretations of the auxiliary operators. The axiom RED follows from the constructions of the $\mathbf{M}$-semantic clauses for D_G, L_{D_G}, I_{D_G}.
- Completeness: Showing completeness for logics with DK operators with respect to the common (possible-worlds) epistemic structures is quite standard in the literature (e.g. Fagin et al. (1992), (Fagin et al., 1995, Chapter 3), Gerbrandy (1999)). In particular, and following (Gerbrandy, 1999, pp. 64-65), the proof requires the construction of *pseudo-models*, whereby distributed knowledge of a group is interpreted through a primitive relation, rather than in terms of the individual relations. Under these models, DK essentially acts as a normal modal operator, therefore soundness and completeness with respect to pseudo-models follows as usual in Modal Logic. Any pseudo-model can be unravelled to a tree-like structure that preserves the truth of all formulas. The unravelled model can then be folded to a proper model, such that each formula that is pseudo-satisfied in the unravelled model is satisfied in the folded model as well.

  The procedure can be straightforwardly applied for showing completeness of Λ_{DK} with respect to epistemic ALMSs, as the latter are essentially possible-worlds, epistemic structures, augmented by syntactic functions. Constructing pseudo-models for ALMSs goes as above; the only difference here is that there are additional syntactic functions: the availability function R and the awareness-like function I_{D_G}, used to interpret the A_j and I_{D_G} operators, respectively. Showing completeness with respect to the pseudo-models, as well as the remaining steps of the proof, are not affected by them, given that in the canonical model, the corresponding functions will be given by the sets $\{\rho \mid A_j\rho \in w\}$ and $\{\phi \mid I_{D_G}\phi \in w\}$, respectively.

□

8.3.3 Dynamic axiomatization

We focus on the behaviour of resolution and inference under epistemic ALMSs. Dynamic formulas are interpreted as indicated by their original clauses, only now at the reduced structures corresponding to M, M_G, and $\mathsf{M} \otimes \mathsf{C}$. More specifically, consider an epistemic RSMM M and its reduced epistemic ALMS $\mathbf{M}$. If an update, of resolution or inference, takes place, then we get an updated M' and thus an updated ALMS $\mathbf{M}'$ corresponding to it. We observe that $\mathbf{M}'$ is such that an updated awareness-like function I'_{D_G} is given in terms of I_{D_G}, i.e. the awareness-like function of $\mathbf{M}$. That is, the new values are set expressions of the original ones. We present the updated functions below, using Definition 8.1.5, Definition 7.2.7, Definition 8.3.1.

- After resolution of G, the awareness-like function for group H is given by:

  ▷ If $G \cap H = \varnothing$:

$$\begin{aligned}
\mathrm{I}'_{D_H}(w) &= \bigcap_{u \in W^I \cap \longrightarrow'_H(w)} V_I(u) \\
&= \bigcap_{u \in W^I \cap \longrightarrow_H(w)} V_I(u) \\
&= \mathrm{I}_{D_H}(w)
\end{aligned}$$

  ▷ If $G \cap H \neq \varnothing$:

$$\begin{aligned}
\mathrm{I}'_{D_H}(w) &= \bigcap_{u \in W^I \cap \longrightarrow'_H(w)} V_I(u) \\
&= \bigcap_{u \in W^I \cap \longrightarrow'_{H \cap G}(w) \cap \longrightarrow'_{H \setminus G}(w)} V_I(u) \\
&= \bigcap_{u \in W^I \cap \longrightarrow_G(w) \cap \longrightarrow_{H \setminus G}(w)} V_I(u) \\
&= \bigcap_{u \in W^I \cap \longrightarrow_{G \cup H}(w)} V_I(u) \\
&= \mathrm{I}_{D_{G \cup H}}(w)
\end{aligned}$$

- After $\mathsf{C}_{\mathsf{ALL}}$, whereby all agents apply a rule ρ, we easily get that:

$$\begin{aligned}\mathrm{I}'_{D_H}(w, e_1) &= \bigcap_{(w',e')\in(W^I)' \cap \longrightarrow'_H(w,e_1)} V'_I(w', e') \\ &= \begin{cases}\mathrm{I}_{D_H}(w) \cup \{con(\rho)\}, \text{ if } (W^I)' \cap \longrightarrow'_H (w, e_1) \neq \varnothing \\ \mathrm{I}_{D_H}(w) \cup \overline{\mathrm{I}_{D_H}(w)}, \text{ if } (W^I)' \cap \longrightarrow'_H (w, e_1) = \varnothing\end{cases}\end{aligned}$$

- After $\mathsf{C}_{\mathsf{SOME}}$, whereby agents in G apply a rule ρ, we get as in Section 7.4.3:

  ▷ Regarding worlds of the updated model generated by e_1:

$$\mathrm{I}'_{D_H}(w, e_1) = \begin{cases}\mathrm{I}_{D_H}(w), \text{ if } (W^I)' \cap \longrightarrow'_H (w, e_1) \neq \varnothing \text{ and } G \cap H = \varnothing \\ \mathrm{I}_{D_H}(w) \cup \{con(\rho)\}, \text{ if } (W^I)' \cap \longrightarrow'_H (w, e_1) \neq \varnothing \text{ and } G \cap H \neq \varnothing \\ \mathrm{I}_{D_H}(w) \cup \overline{\mathrm{I}_{D_H}(w)}, \text{ if } (W^I)' \cap \longrightarrow'_H (w, e_1) = \varnothing\end{cases}$$

  ▷ Regarding worlds of the updated model generated by e_0:

$$\mathrm{I}'_{D_H}(w, e_0) = \mathrm{I}_{D_H}(w)$$

We now proceed as we did in Section 7.4.3, where we explained how reduction axioms can be obtained in cases when the new sets obtained through an update of a syntactic function cannot be described by means of the static language alone. We now follow this procedure, but tailored to the function I_{D_G}. This is because, as shown above, the updated values are too given in terms of the original ones, reflecting the refinement induced by each action.

8.3.5. Definition (Language $\mathcal{L}_{\mathsf{DK}_{\mathsf{SE}}}$). The formulas ϕ and the set expressions over formulas Ω of the language $\mathcal{L}_{\mathsf{DK}_{\mathsf{SE}}}$ are given by:

$$\begin{gathered}\phi ::= p \mid z_1 s_1 + \ldots + z_n s_n \geq z \mid \neg\phi \mid \phi \wedge \psi \mid A_j\rho \mid D_G\phi \mid L_{D_G}\phi \mid [\Omega]\phi \\ \Omega ::= I_{D_G} \mid \{\phi\} \mid \overline{\Omega} \mid \Omega \cup \Omega\end{gathered}$$

The I_{D_G} operators are now re-expressed as set-expression operators of the form $[\Omega]$. This allows us to capture syntactically the sets generated by actions of resolution and inference.

8.3.6. Definition (Truth clauses for new formulas). Given ALMS $\mathbf{M}$ and $w \in$ W, the truth clauses for the new formulas are:

$$\begin{array}{ll}\mathbf{M}, w \models [I_{D_G}]\psi \text{ iff } \psi \in \mathrm{I}_{D_G}(w) & \mathbf{M}, w \models [\{\phi_1\}]\phi_2 \text{ iff } \phi_1 = \phi_2 \\ \mathbf{M}, w \models [\overline{\Omega}]\phi \text{ iff } \phi \notin \Omega & \mathbf{M}, w \models [\Omega_1 \cup \Omega_2]\phi \text{ iff } \phi \in (\Omega_1 \cup \Omega_2)\end{array}$$

8.3.7. Theorem (Axiomatization for extended language $\mathcal{L}_{\mathsf{DK}_{\mathsf{SE}}}$). *The system given by the axioms of Table 8.2, alongside Definition 8.3.3, is sound and complete with respect to epistemic ALMSs.*

$[\{\phi\}]\phi$	$\neg[\{\phi_1\}]\phi_2$, for $\phi_1 \neq \phi_2$
$[\overline{\Omega}]\phi \leftrightarrow \neg[\Omega]\phi$	$[\Omega_1 \cup \Omega_2]\phi \leftrightarrow ([\Omega_1]\phi \vee [\Omega_2]\phi)$

Table 8.2: The axioms for set-expression operators

The set expressions allow us to express the effect of actions of resolution and inference, e.g. of those manifested in the case studies. This is because, as shown before, both types of actions induce set-expression definable updates, in that the updated values of I'_{D_G} are set expressions in terms of values of the original I_{D_G} values. We will denote the set expressions yielding the new values following event e of action model C by $[I_{D_G}^{(\mathsf{C},e)}]$.

To give reduction axioms for resolution, we need to express the updated values of capacity and costs in the language. To that end, take:

$$cp_j^G := \begin{cases} cp_j - c_G, & \text{when } j \in G \\ cp_j - c_{G_1}, \text{otherwise} \end{cases}$$

while $c_\rho^G := c_\rho$ (also $c_{\rho_1}^G := c_{\rho_1}$) and $c_H^G := c_H$ (also $c_{H_1}^G := c_{H_1}$).

To give reduction axioms for actions of inference, we need to express the updated values of capacity and costs in the language as well:

$$cp_j^{(\mathsf{C},e)} := \begin{cases} cp_j - c_\rho, & \text{when } lab_1(e) = \rho \text{ for some } \rho \in \mathcal{L}_R \text{ and } j \in lab_2(e) \\ cp_j - c_{\rho_1}, & \text{when } lab_1(e) = \rho \text{ for some } \rho \in \mathcal{L}_R \text{ and } j \notin lab_2(e) \\ cp_j, \text{otherwise} \end{cases}$$

while $c_\rho^{(\mathsf{C},e)} := c_\rho$ (also $c_{\rho_1}^{(\mathsf{C},e)} := c_{\rho_1}$) and $c_H^{(\mathsf{C},e)} := c_H$ (also $c_{H_1}^{(\mathsf{C},e)} := c_{H_1}$).

We also use $pre(\mathsf{R}_G) := \bigwedge_{i \in G} (cp_i \geq c_G)$. We are now ready to provide reduction axioms for $\langle \mathsf{R}_G \rangle$- and $\langle \mathsf{C}, e \rangle$-operators.

$\langle \mathsf{R}_G \rangle (z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow pre(\mathsf{R}_G) \wedge (z_1 s_1^G + \ldots + z_n s_n^G \geq z)$	
$\langle \mathsf{R}_G \rangle p \leftrightarrow pre(\mathsf{R}_G) \wedge p$	$\langle \mathsf{R}_G \rangle (\phi \wedge \psi) \leftrightarrow \langle \mathsf{R}_G \rangle \phi \wedge \langle \mathsf{R}_G \rangle \psi$
$\langle \mathsf{R}_G \rangle \neg\phi \leftrightarrow pre(\mathsf{R}_G) \wedge \neg\langle \mathsf{R}_G \rangle \phi$	$\langle \mathsf{R}_G \rangle A_j \rho \leftrightarrow pre(\mathsf{R}_G) \wedge A_j \rho$
$\langle \mathsf{R}_G \rangle L_{D_H} \phi \leftrightarrow pre(\mathsf{R}_G) \wedge L_{D_{G \cup H}} \langle \mathsf{R}_G \rangle \phi$, if $G \cap H \neq \varnothing$	$\langle \mathsf{R}_G \rangle [I_{D_H}] \phi \leftrightarrow pre(\mathsf{R}_G) \wedge [I_{D_{G \cup H}}] \phi$, if $G \cap H \neq \varnothing$
$\langle \mathsf{R}_G \rangle L_{D_H} \phi \leftrightarrow pre(\mathsf{R}_G) \wedge L_{D_H} \langle \mathsf{R}_G \rangle \phi$, if $G \cap H = \varnothing$	$\langle \mathsf{R}_G \rangle [I_{D_H}] \phi \leftrightarrow pre(\mathsf{R}_G) \wedge [I_{D_H}] \phi$, if $G \cap H = \varnothing$
$\langle \mathsf{R}_G \rangle [\{\phi\}] \phi \leftrightarrow pre(\mathsf{R}_G) \wedge \top$	$\langle \mathsf{R}_G \rangle [\{\phi_1\}] \phi_2 \leftrightarrow pre(\mathsf{R}_G) \wedge \bot$, for $\phi_1 \neq \phi_2$
$\langle \mathsf{R}_G \rangle [\overline{\Omega}] \phi \leftrightarrow \langle \mathsf{R}_G \rangle \neg [\Omega] \phi$	$\langle \mathsf{R}_G \rangle \ [\Omega_1 \cup \Omega_2] \phi \leftrightarrow \langle \mathsf{R}_G \rangle ([\Omega_1] \phi \vee [\Omega_2] \phi)$

Table 8.3: The reduction axioms for resolution

$\langle \mathsf{C}, e\rangle(z_1 s_1 + \ldots + z_n s_n \geq z) \leftrightarrow \mathit{pre}(e) \wedge (z_1 s^{(\mathsf{C},e)} + \ldots + z_n s_n^{(\mathsf{C},e)}) \geq z$	
$\langle \mathsf{C}, e\rangle p \leftrightarrow \mathit{pre}(e) \wedge p$	$\langle \mathsf{C}, e\rangle \neg\phi \leftrightarrow \mathit{pre}(e) \wedge \neg\langle \mathsf{C}, e\rangle\phi$
$\langle \mathsf{C}, e\rangle(\phi \wedge \psi) \leftrightarrow \langle \mathsf{C}, e\rangle\phi \wedge \langle \mathsf{C}, e\rangle\psi$	$\langle \mathsf{C}, e\rangle A_j\rho \leftrightarrow \mathit{pre}(e) \wedge A_j\rho$
$\langle \mathsf{C}, e\rangle L_{D_G}\phi \leftrightarrow \mathit{pre}(e) \wedge \bigwedge_{e \twoheadrightarrow_j e'} L_{D_G}[\mathsf{C}, e']\phi$	$\langle \mathsf{C}, e\rangle[I_{D_G}]\phi \leftrightarrow \mathit{pre}(e) \wedge [I_{D_G}^{(\mathsf{C},e)}]\phi$
$\langle \mathsf{C}, e\rangle[\{\phi\}]\phi \leftrightarrow \mathit{pre}(e) \wedge \top$	$\langle \mathsf{C}, e\rangle[\{\phi_1\}]\phi_2 \leftrightarrow \mathit{pre}(e) \wedge \bot$, for $\phi_1 \neq \phi_2$
$\langle \mathsf{C}, e\rangle[\overline{\Omega}]\phi \leftrightarrow \langle \mathsf{C}, e\rangle\neg[\Omega]\phi$	$\langle \mathsf{C}, e\rangle[\Omega_1 \cup \Omega_2]\phi \leftrightarrow \langle \mathsf{C}, e\rangle([\Omega_1]\phi \vee [\Omega_2]\phi)$

Table 8.4: The reduction axioms for inference

The reduction axioms for the Boolean cases, rule-availability, and set expressions are straightforward to read. The reduction axiom for inequalities is such to reflect, with the help of our abbreviations, the resource consumption that each action induces. The reduction axioms for L_{D_G} operators is in DEL-lines, as L_{D_G}-operators are possible-world quantifying operators that behave as in standard DEL. Under resolution, they behave as D_G does in frameworks with similar communicative actions (Ågotnes and Wáng, 2017; Baltag and Smets, 2020). Under product updates, they behave as normal operators do (van Ditmarsch et al., 2007). The crucial part is the reduction axioms for I_{D_G} which capture exactly the fact that awareness-like function is updated in a principled way, expressable in terms of the original values. The axiom for D_G can be derived from those for L_{D_G} and I_{D_G}.

8.3.8. Proposition (Reduction axioms). *The reduction axioms for resolution (Table 8.3) and for inference (Table 8.4) are valid in epistemic ALMSs.*

Proof:
The claim is easy to check for the axioms reducing dynamic formulas involving the Boolean connectives, linear inequalities, and set-expression operators. We will focus on the cases for L_{D_H} and I_{D_H}. Consider the reduction axioms for resolution:

- Let M be arbitrary epistemic RSMM, w an arbitrary possible world of it and assume $\mathbf{M}_{\mathsf{M}}, w \models \langle \mathsf{R}_G\rangle L_{D_H}\phi$. This is the case iff $\mathbf{M}_{\mathsf{M}}, w \models \mathit{pre}(\mathsf{R}_G)$ and $\mathbf{M}_{\mathsf{M}_G}, w \models L_{D_H}\phi$. The latter amounts to $\mathbf{M}_{\mathsf{M}_G}, u \models \phi$ for all $u \in W^P$ such that $w \longrightarrow^{r'}_H u$, where $\longrightarrow^{r'}_H := \cap_{j\in H} \longrightarrow^{r'}_j$. Using Definition 8.1.5 and Definition 8.3.1:
  - ▷ If $G \cap H = \varnothing$, then $\longrightarrow^{r'}_H = \cap_{j\in H} \longrightarrow^{r}_j$. Therefore, $\mathbf{M}_{\mathsf{M}_G}, u \models \phi$ for all $u \in W^P$ such that $w \longrightarrow^{r}_H u$. This, in combination with $\mathbf{M}_{\mathsf{M}}, w \models \mathit{pre}(\mathsf{R}_G)$ (shown before), ensures that $\mathbf{M}_{\mathsf{M}}, w \models \mathit{pre}(\mathsf{R}_G) \wedge L_{D_H}\langle \mathsf{R}_G\rangle\phi$.
  - ▷ If $G \cap H \neq \varnothing$, then $\longrightarrow^{r'}_H = \bigcap_{j\in G\cap H} \longrightarrow^{r'}_j \cap \bigcap_{j\in H\setminus G} \longrightarrow^{r}_j = \bigcap_{j\in G} \longrightarrow^{r}_j \cap \bigcap_{j\in H\setminus G} \longrightarrow^{r}_j = \longrightarrow^{r}_G \cap \longrightarrow^{r}_{H\setminus G} = \longrightarrow^{r}_{G\cup H}$. Therefore, $\mathbf{M}_{\mathsf{M}_G}, u \models \phi$ for all $u \in W^P$ such that $w \longrightarrow^{r}_{G\cup H} u$. This, in combination with $\mathbf{M}_{\mathsf{M}}, w \models \mathit{pre}(\mathsf{R}_G)$ (shown before), ensures that $\mathbf{M}_{\mathsf{M}}, w \models \mathit{pre}(\mathsf{R}_G) \wedge L_{D_{G\cup H}}\langle \mathsf{R}_G\rangle\phi$.

The converse directions follow analogously.

- Let M be arbitrary epistemic RSMM, w an arbitrary possible world of it and assume $\mathbf{M}_{\mathsf{M}}, w \models \langle \mathsf{R}_G \rangle [I_{D_H}]\phi$. This is the case iff $\mathbf{M}_{\mathsf{M}}, w \models pre(\mathsf{R}_G)$ and $\phi \in \mathrm{I}'_{D_H}(w)$. Using the computed values of the updated functions:

  ▷ If $G \cap H = \varnothing$, then $\phi \in \mathrm{I}'_{D_H}(w)$ iff $\phi \in \mathrm{I}_{D_H}(w)$, i.e. $\mathbf{M}_{\mathsf{M}}, w \models [I_{D_H}]\phi$. As a result, $\mathbf{M}_{\mathsf{M}}, w \models pre(\mathsf{R}_G) \wedge [I_{D_H}]\phi$ as desired.

  ▷ If $G \cap H \neq \varnothing$, then $\phi \in \mathrm{I}'_{D_H}(w)$ iff $\phi \in \mathrm{I}_{D_{G \cup H}}(w)$, i.e. $\mathbf{M}_{\mathsf{M}}, w \models [I_{D_{G\cup H}}]\phi$. As a result, $\mathbf{M}_{\mathsf{M}}, w \models pre(\mathsf{R}_G) \wedge [I_{D_{G\cup H}}]\phi$ as desired.

  The converse directions follow analogously.

The axioms for inferential actions are obtained in a similar fashion. For L_{D_G} operators, the reduction axiom is obtained along the lines of the reduction axiom for standard (distributed) knowledge in DEL (van Ditmarsch et al., 2007). For I_{D_G} operators, the reduction axiom makes use of the computed values of the updated functions, as in the reduction axioms for resolution. □

8.3.9. Theorem (Full axiomatization). *The axioms and rules of Definition 8.3.3, Table 8.2, Table 8.3, Table 8.4 and the Action Necessitation rules (from ϕ, infer $[\mathsf{R}_G]\phi$, and from ϕ infer $[\mathsf{C}, e]\phi$) provide the full sound and complete system for the language $\mathcal{L}_{\mathsf{DK_{SE}}}$ extended by the dynamic operators with respect to epistemic ALMSs and the actions of resolution and inference.*

Proof:
The result follows from Theorem 8.3.4, Theorem 8.3.7 and Proposition 8.3.8. □

8.4 Conclusions and further work

The objections against the EL modelling of unbounded agents have repercussions for group reasoning as well. The notion of DK is instrumental in illustrating this, because it presupposes that agents can undertake unlimited actions of communication and inference. We looked into actualizations of DK under bounded resources, using epistemic RSMMs and actions for communication and inference. We furthermore showed that the techniques underlying the static and dynamic axiomatization from Section 7.4 still apply in a natural way.

This approach addresses, in effect, the manifestation of the problem of logical omniscience in group settings, taking into account experimental results and philosophical proposals (in the lines of the alternative picture), much like Chapter 7 did for multi-agent doxastic reasoning. Departing from this problem, this approach also demarcates the communicative and inferential actions underlying whether and how DK is actualized. As van Benthem (2008c) argues, information goes hand in hand with the processes that create, modify, and convey it; this analysis naturally applies to deliberating groups, and importantly, to resource-bounded ones.

One direction for future work concerns non-ideal higher-order reasoning, and hence connections between DK and CK. As with deductive reasoning, this requires the introduction of effortful steps (e.g. for introspection and reasoning about other agents) and the use of experimental results showing that groups usually act on a large, but finite, degree of mutual knowledge as if they had CK (*almost common knowledge* (Rubinstein, 1989)). It can therefore be pursued through a combination of the results of this chapter and of the previous one. Clark and Marshall (1981) delineate heuristics according to which agents act as if they had CK. In order to embed such heuristics in our model we might need a richer logical system that combines System 1 and System 2 processes (Chapter 5).

Another direction is towards modelling more complex scenarios and variants of psychological tasks. One way to do so is through accounting for a greater repertoire of resolution-like actions. This direction was explained in Section 8.2.4. Alternatively, we can account for other actions that may be taking place in reasoning tasks. For example, the subjects of selection tasks often choose to turn card 4 as well – according to some theorists, because of the matching bias (Section 2.3). This may also be accounted for in combination with frameworks including different actions run by System 1 and System 2 actions (Chapter 5).

On another note, group reasoning, in this attempt, can be better than individual in ways that agree with the view that at the upper limit groups perform as their best member (Laughlin and Ellis, 1986), and the distribution of skills observed by Geil (1998) and Trognon et al. (2011). However, these studies also emphasize the facilitative effect of *dialogue* in group performance. The authors indicate that the group environment encouraged agents to raise different alternatives and challenge each other, which in turn lends support to the dialogical basis of rationality. To account for these facilitation effects, we need a dialogical/inquisitive system, where raising alternatives is explicitly modelled, showing why in multi-agent settings, the application of rules is easier than in single-agent ones.

Chapter 9

Future work and closing remarks

Wrapping up, we discuss two broader issues that could follow the questions addressed in the previous chapters and build bridges to other research agendas. We finally provide some closing remarks and reflections on the thesis as a whole.

9.1 Notes on future work

9.1.1 On probabilistic reasoning

A crucial aspect of human reasoning that has not been incorporated in the foregoing frameworks is *probabilistic reasoning*. The question of how people reason under uncertainly has attracted the attention of formal modellers and empirical scientists alike. The resulting debates have unfolded in parallel to the ones surveyed in Chapter 2. We believe that the argumentation pattern adopted throughout that chapter applies to questions of probabilistic reasoning as well. More specifically, the principles of classical probability theory, have been considered as constituents, next to those of classical logic, of the standard rationality norms.[1] Much like the logical aspect, the probabilistic has also been challenged in the face of psychological evidence. One of the well-known illustrations of this is the so-called *Linda problem* (Tversky and Kahneman, 1983, p.297):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

The subjects are then asked to evaluate which is the more probable statement on basis of the description of this person:

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

[1] Also, see Halpern (2017) for rationality postulates when reasoning under uncertainty.

The vast majority of subjects chose option 2. However, this choice goes against a rule of classical probability, the *conjunction rule*, according to which the probability of two events occurring together is always less than or equal to the probability of either one of them occurring individually.

In short, we observe the same gap between the normative principles and the principles that appear to be included in our reasoning competence. Yet again, a challenge for SRT arises, this time concerning its probabilistic aspect.

In our view, the discussion (counterarguments defending SRT and criticisms thereof) and the conclusions ensuing from this challenge can be seen on a par with those in Section 2.4. That is, contrary to attempts of defending SRT and its alignment with classical probability norms, we advocate for an alternative picture. This takes seriously into account the demonstrated limitations of human agents and what possibly gives rise to violations of the standard norms. This is to ensure that *ought* does imply *can*, also with respect to probabilistic reasoning.

But what about the *formal* study of an alternative picture that encompasses probabilistic reasoning as well? This hinges on two issues: first, the very combination of the formal study of logic and probability, and second, the question on whether this combination can be made alternative picture-friendly – and, of particular interest to our attempt, whether the combination can be successfully paired with the methods pursued to attain a formalization of the alternative picture for the logical aspect.

Considering the first issue, the idea of logic and probability theory as joint forces in the formal study of human reasoning might at first sound unnatural, for the former is occupied with matters in the realm of certainty while the latter is occupied with matters in the realm of uncertainty (Hájek, 2001; Demey et al., 2019). Moreover, the approach of logic to inference is qualitative while the approach of probability to inference is quantitative. Contrary to that, many authors have contributed towards resolving this misconception, often addressing a range of interesting epistemological questions along the way (see, for example, van Benthem (2017) and Leitgeb (2017)).

Considering the second issue, we examine how to apply the resource-sensitive view advocated throughout the previous chapters to the interface of logic and probability. In what follows, we sketch two ways to do so.

- In Chapter 3, we discussed the work of Bjerring and Skipper (2018), and its relationship to ours. Recall that this work is a member of the family of approaches using impossible-worlds semantics against logical omniscience, and most importantly, one of the closest to our proposal, for it too uses reasoning-oriented dynamic actions to strike the balance between logical omniscience and logical incompetence. Skipper and Bjerring (2020) have applied this idea to a probabilistic setting as well, in order to avoid idealization but still respect the idea that rational degrees of belief are constrained by laws of probability. We expect that our cost-monitoring reasoning processes as well as the technical results of Part II can also be similarly attuned to a probabilistic setting.

- The questions of Chapter 4 and Chapter 5 were explored through the use of plausibility models. These may assist in building a bridge to a probabilistic, resource-sensitive framework. Plausibility orderings make for a qualitative representation of belief entrenchment and dispositions to belief revision. However, the DEL framework can also be extended to a quantitative setting, representing degrees of belief and embedding probabilistic insights. Baltag and Smets (2008a) combine DEL, the AGM belief revision theory (as interpreted on plausibility models), and the Popper–Reyni–de Finetti extension of Bayesian probabilistic conditionalization. Among others, they introduce *discrete conditional probabilistic models for knowledge and conditional belief* and discuss their relationship with (standard) plausibility models. The outcome is a correspondence between the two types of models, that allows us to obtain a plausibility model when given a probabilistic one and vice-versa. It therefore makes sense to ask whether this representation could be extended to their resource-sensitive counterparts. Establishing an analogous result makes for a suitable basis for an extension of the logics of the previous chapters to a probabilistic setting.

The aforementioned directions are only indicative of the ways of encompassing probabilistic reasoning in a formal analysis that takes resource-bounds into account (see, for example, Icard (2014) and Nguyen and Rakib (2019), for more). While these questions raise enough philosophical and technical challenges to constitute a project on their own, with these lines, we wish to endorse a view that does not look at logic and probability as divorced topics.

9.1.2 On higher-order reasoning in social networks

The formal study of social networks and of phenomena especially emerging in their environment is increasingly popular and promising. In this section, we argue for the connection between this study and the observations of Part I and Part III.[2]

More specifically, in diverse social phenomena, the way a social network shapes individual opinions depends on the information the individuals in the network have about one another. This is clear in situations such as:

- *Pluralistic Ignorance*, where the majority of the individuals do not accept an idea, and yet they go along with it because they assume, incorrectly, that everybody else accepts it (Katz and Allport, 1931).
- *Opinion Diffusion*, where agents adopt certain ideas influenced by their perception on whether the idea has been accepted by their peers (Easley and Kleinberg, 2010).

[2] The section is partly based on Solaki and Velázquez-Quesada (2020).

Formalizations of these phenomena (e.g. respectively, Rendsvig (2014); Proietti and Olsson (2014) and Baltag et al. (2019)) rely on higher-order reasoning. For example, the latter uses *threshold modelling*: according to it, agents adopt a new behaviour/product/opinion, when (they know that) the proportion of their neighbors who have already adopted it meets a given threshold. The underlying logical tool to address the epistemic dimension of these phenomena is based on standard EL with its relational semantics. This is also the case with formalizations of the Bystander Effect (Rendsvig, 2014) and of the creation and evolution of networks (Seligman et al., 2013; Smets and Velázquez-Quesada, 2017). As a result, the formalizations inherit the objections presented in Chapter 2 and Chapter 6. That is, these proposals make use of logical tools that make strong idealizations about the involved agents, not only with respect to their reasoning abilities, but also with respect to their competence in identifying what other people believe or prefer.

In particular, an important feature of how people function in social networks, that is often overlooked, is that of ToM. Theory of Mind is important because our behaviour in social networks relies on reasoning about ourselves and others. However, we have seen that this ability is not unbounded nor does it come without any cognitive cost. For example, specifically concerning social networks, there is evidence that ToM affects (sets an upper limit to) one's innermost network layer (Stiller and Dunbar, 2007). Thus (a) reasoning about others is evidently involved in the formation and evolution of social networks, and (b) there is experimental evidence on limits in our ability to reason about others in our social network. Therefore, the formal study of social networks should be intertwined with the formal study of social cognition.

In Chapter 6, after explaining why EL might not be well-suited for representing realistic higher-order attributions, we proposed a system that gives an alternative to the traditional logical modelling of belief ascriptions, which is closer to the workings of belief formation. Its most important feature was the contrast between a "simple" model for visibility and communication (including misinformation and lying, which are common in social networks), and a "complex" clause for interpreting mental state attributions (depending on agents' recollection of visibility and communication facts). This can help us track why in certain cases these tasks become too cognitively difficult, in agreement with experimental findings.

We expect that our framework for belief attributions can provide a more realistic basis for the epistemic/doxastic dimension of many social phenomena.

- Regarding the modelling of opinion diffusion: by asking that agents adopt x when the proportion of their neighbors that *they believe have adopted* x meets a threshold; this is a higher-order ascription, and could be interpreted as in the alternative (not relational) semantics.

- Regarding the modelling of network creation: agents form connections with the agents they believe they are in agreement with; however, the comparison between their own beliefs and the beliefs of others could be grounded on the alternative semantics.

- The effect of ToM skills and memory capacity on the formation of social networks can be accounted for by limiting the number of states an agent is able to recall, or the number of agents whose "visibility" she can keep track of.

Similar modifications may apply to phenomena like Pluralistic Ignorance and Bystander Effect, whose formalization also involves belief attributions to others. These are some initial examples for informing the formal logical study of social network phenomena with insights from the empirical study of social cognition.

9.2 Closing remarks

Closing the thesis, we recapitulate the contributions of each chapter and reflect on their adequacy towards fulfilling our desiderata.

In Chapter 2, we argued for an alternative rationality picture, as a response to the challenges facing the standard one, due to empirical findings on human deductive reasoning, introspection, and reasoning about others. According to the alternative picture, the design of the norms for human reasoning should be informed not only by our intuitions or our general philosophical theories, but also by empirical facts on the limits of cognition and on the variety of processes that give rise to its outputs. We explained why this shift necessitates a shift in the epistemic logical modelling, to the mutual benefit of disciplines studying human reasoning. We then embarked on the task of providing alternative logical frameworks to this end.

We started off with Chapter 3, modelling explicitly the evolution of a single agent's knowledge via steps of deductive reasoning. This was achieved through the introduction of a dynamic, resource-sensitive, impossible-worlds semantics. By using impossible worlds, we avoided the extreme of logical omniscience. By using applications of inference rules that gradually refine a fallible agent's epistemic state, we avoided the extreme of logical incompetence. Importantly, these applications can be undertaken only to the extent allowed by cognitive resources, thereby specifying exactly what draws the cutoff on reasoning processes. The dynamic dimension therefore acts as the bridge between the logical formalization and the empirically indicated difficulty of different deductive steps. Moreover, we established a connection between our impossible-worlds structures and syntactic, awareness-like ones. We showed that the connection is instrumental towards providing a sound and complete axiomatization. As a result, there exists a way to combine the benefits of dynamic impossible-worlds semantics in modelling non-idealized agents and the use of common Modal Logic techniques.

The resource-sensitive modelling goes beyond the strong notion of knowledge. Besides, the alternative rationality view asks for more than a coarse-grained classification of mental states. Contrary to misconceptions on the relationship of logic with epistemology and cognitive science, DEL has the tools to overcome this obstacle. In Chapter 4, we designed suitable plausibility models that pave the way towards bringing more nuanced attitudes under the resource-sensitive roof. The models were shown to be reducible to syntactic plausibility structures, enriching the correspondence picture, but also reaping its technical benefits. Because of the contribution of plausibility modelling in the study of diverse interaction dynamics, we could also model the interplay between external information (received from sources of varying reliability – be it with or without effort) and the internal reasoning steps of the agent. The resource-sensitive modelling was therefore proven to be flexible enough to embed the established progress of DEL, represent interaction and inference together, and thus draw more connections with the alternative picture.

Apart from empirical facts on the limits of deductive reasoning, the alternative picture also incorporates evidence on the dual process theories of reasoning. The setting of Chapter 5 built on the resource-sensitive plausibility modelling to do justice to the effect of both System 1 and System 2 processes in the evolution of an agent's reasoning. Via this, we managed to approach reasoning scenarios that are usually neglected in the logical literature. We finally delved deeper into the understanding of the "*ought implies can*" principle (when used for the logician's purposes), by taking into account that reasoning can be individuated by different types of processes, over which we might not always exercise deliberate control.

We then moved to multi-agent reasoning. In Chapter 6, we focused on the *formation* of mental state attributions. In particular, we introduced a dynamic temporal setting and an alternative semantic interpretation that relies on each agent's memory, visibility, and communication, about facts and other agents. The setting was applied to FBTs, well-known tasks studying people's ToM. We discussed its technical features, as well those connecting it to elements of the alternative picture. We also evaluated the setting against two criteria of robustness and faithfulness, which are of particular interest to the project of formalizing FBTs.

Following the formation of mental state attributions, we aimed at the *manipulations* occurring in multi-agent reasoning (Chapter 7). These include both actions of deductive reasoning and actions of higher-order reasoning, e.g. of introspection and of attribution of rule-applications to other agents. Their uniform treatment was achieved through a combination of the resource-sensitive semantics with a novel type of action models. As a result, we could extend our contributions to higher-order reasoning and ensure that the *ought implies can* principle is respected in a multi-agent framework as well. On the more technical side, we developed the correspondence picture to encompass multi-agent structures and utilized the result, along with an alternative method to providing reduction axioms, to obtain a sound and complete axiomatization.

In Chapter 8, we targeted bounded group reasoning, and in particular, its manifestation in actualizing a group's distributed knowledge. The notion, in its traditional understanding, neglects the fact that cognitive capacities of group members interfere with what can be actually achieved by it. The challenges posed by the alternative picture for the logical treatment of an individual agent's reasoning appear in the collective level too. We therefore developed the multi-agent framework of Chapter 7 to capture group reasoning as well. We brought together our action models with effortful communicative actions, whereby members share their knowledge within the group. We applied the framework to group reasoning tasks and identified reasons why group reasoning might be better than individual. We finally extended the correspondence and witnessed its applicability to logics that include group knowledge operators.

While the workings of human reasoning certainly go beyond the aspects hitherto addressed, these frameworks were designed to act as formal counterparts of elements of the alternative picture, with the hope of advancing the exchange between EL and other disciplines studying human reasoning.

The project was fueled by a nuanced view in revisiting the rationality norms, which has repercussions for epistemology and logic. The bottom line has been, not to reject the normative purposes of the logical modelling, but rather to aim at a *can*-implying *ought*, as our guiding principle. In the following lines, we reflect on how the thesis as a whole has dealt with this guiding principle.

The extreme of the traditional *ought* has been avoided due to our modifications of possible-worlds semantics. However, the common objection against this strategy, especially fired against impossible-worlds semantics, is that it delivers agents who exhibit entirely erratic behaviour; the danger that then lurks is that one does not formalize an *ought* that implies *can*, but rather an *ought* that is derived by *is* (and even worse, the *is* of a logically incompetent agent). However, we did argue against a nihilistic view as a replacement of the standard rationality picture. Thus, any modification of the possible-worlds apparatus that merely formalizes a rationality picture we have already rejected would not do.

This is where the dynamics enter the picture. While we allowed for agents who fall short of the expectations of EL, we accounted both for the intricate processes that refine their epistemic/doxastic states, and for limits of these processes. Still, logics diverging from the received view are sometimes accused for being too weak to deliver any non-trivial, interesting, and illuminating insights on our mental processes. Against this, we argue that, due to the very nature of human reasoning *processes*, these insights lie on the *dynamic* dimension of logical modelling.

This is why we have transferred the "burden of proof" for showing that we deliver the intended, balanced notion of *ought* to our reasoning actions. Shifting the task of revealing interesting features of reasoning processes to the dynamics is a very much deliberate choice, opposed to forcing unrealistically strong logical principles, that operate on an "all-or-nothing" basis (the agents are either omniscient, fully introspective, and reason about one another at any modal depth, or they are incompetent and erratic).

In more detail: an isolated, static "snapshot" of an agent of a resource-sensitive model might indeed capture her in a fallible state, e.g. entertaining an incomplete or inconsistent world, and would not alone reveal many illuminating insights about her. But a snapshot of a real reasoner's mental state would also say very little on the wealth of the person's mental trajectory. That is, the logical formalizations rely on dynamics to shed light on our cognitive lives, because our cognitive lives are themselves shaped by interconnected processes and not by static snapshots in which we are either perfect or utterly crazy reasoners. To use fallibility (and its formal witnesses in the logical model) to discard one's reasoning processes (and their formal modelling) altogether is guilty of the nihilism that proponents of the traditional "ought" would warn us against. The truly non-trivial and interesting insights can be revealed not by fussing about our, perhaps unfortunate, snapshots, but by tracking our evolving enterprise to refine our mental states in their aftermath.

Consider the process of writing a PhD thesis. When PhD students start their projects, they clearly have not unveiled every true statement in their research field (why else would they even embark on the project?). They may have many open questions, believe in false conjectures, be biased towards one or another school of thought, forget a detail they once read or were told about, fail to reason accurately about the beliefs of their collaborators, and so on. But facing this, should we deem the reasoners to be worthless and withdraw to an absolutist demand that researchers ought to know the outcome of their research, even before conducting it? Contrary to this, we argue that what is truly interesting lies somewhere else: in that, *despite* all the above, the students break down big questions to smaller ones, put effort in gradually resolving them, communicate with peers to learn more, confront their own biases, recharge and start again, recollect useful information they once heard in a conference, acknowledge the expertise of their teammates, and so on. We *can* extrapolate illuminating insights from this process; to accept this certainly does not amount to submitting to nihilism.

In a very different context, Emma Goldman once declared "*What I believe* is a process rather than a finality. Finalities are for gods and governments, not for the human intellect". The frameworks of this thesis, and in fact the very process of writing it, can only echo this statement.

Bibliography

Addis, T. and Gooding, D. (1999). Learning as collective belief-revision: Simulating reasoning about disparate phenomena. In *AISB Symposium*, pages 6–9.

Albore, A., Alechina, N., Bertoli, P., Ghidini, C., Logan, B., and Serafini, L. (2006). Model-checking memory requirements of resource-bounded reasoners. In *AAAI*, volume 6, pages 213–218.

Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530.

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

Alechina, N., Logan, B., Nga, N. H., and Rakib, A. (2009a). A logic for coalitions with bounded resources. In *IJCAI*, pages 659–664.

Alechina, N., Logan, B., Nga, N. H., and Rakib, A. (2009b). Verifying time, memory and communication bounds in systems of reasoning agents. *Synthese*, 169(2):385–403.

Alechina, N., Logan, B., Nguyen, H. N., and Rakib, A. (2008). Reasoning about other agents' beliefs under bounded resources. In *International Workshop on Knowledge Representation for Agents and Multi-Agent Systems*, pages 1–15. Springer.

Apperly, I. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., and Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, 17(10):841–844. PMID: 17100782.

Areces, C., Fervari, R., and Hoffmann, G. (2015). Relation-changing modal operators. *Logic Journal of the IGPL*, 23(4):601–627.

Arslan, B., Taatgen, N., and Verbrugge, R. (2013). Modeling developmental transitions in reasoning about false beliefs of others. In *Proceedings of the 12th International Conference on Cognitive Modeling, Ottawa: Carleton University*, pages 77–82.

Artemov, S. and Fitting, M. (2020). Justification Logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition.

Aucher, G., van Benthem, J., and Grossi, D. (2018). Modal logics of sabotage revisited. *J. Log. Comput.*, 28(2):269–303.

Bago, B. and Neys, W. D. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, 158:90 – 109.

Balbiani, P., Fernández-Duque, D., and Lorini, E. (2019). The dynamics of epistemic attitudes in resource-bounded agents. *Studia Logica*, 107(3).

Ball, L. J. and Thompson, V. A. (2018). Belief bias and reasoning. In Ball, L. and Thompson, V., editors, *The Routledge International Handbook of Thinking and Reasoning*, pages 16–35. Routledge.

Baltag, A., Christoff, Z., Rendsvig, R. K., and Smets, S. (2019). Dynamic epistemic logics of diffusion and prediction in social networks. *Studia Logica*, 107:489–531.

Baltag, A., Moss, L. S., and 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, pages 43–56, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Baltag, A., Özgün, A., and Sandoval, A. L. V. (2018). APAL with memory is better. In Moss, L. S., de Queiroz, R. J. G. B., and Martínez, M., editors, *25th International Workshop WoLLIC 2018*, volume 10944 of *Lecture Notes in Computer Science*, pages 106–129. Springer.

Baltag, A. and Renne, B. (2016). Dynamic epistemic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Baltag, A., Renne, B., and Smets, S. (2014). The logic of justified belief, explicit knowledge, and conclusive evidence. *Annals of Pure and Applied Logic*, 165(1):49 – 81. The Constructive in Logic and Applications.

Baltag, A. and Smets, S. (2008a). Probabilistic dynamic belief revision. *Synthese*, 165(2):179–202.

Baltag, A. and Smets, S. (2008b). 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. and Smets, S. (2009). Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 41–50.

Baltag, A. and Smets, S. (2011). Keep changing your beliefs, aiming for the truth. *Erkenntnis*, 75(2):255–270.

Baltag, A. and Smets, S. (2013). Protocols for belief merge: Reaching agreement via communication. *Logic Journal of the IGPL*, 21(3):468–487.

Baltag, A. and Smets, S. (2020). Learning what others know. In *LPAR*, pages 90–119.

Bara, B. G., Bucciarelli, M., and Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. *The American Journal of Psychology*, 108(2):157–193.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a “theory of mind” ? *Cognition*, 21(1):37 – 46.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes.* MIT Press.

Ben-David, S. and Ben-Eliyahu-Zohary, R. (2000). A modal logic for subjective default reasoning. *Artificial Intelligence*, 116(1-2):217–236.

Berto, F. and Jago, M. (2019). *Impossible worlds.* Oxford University Press.

Bezhanishvili, N. and Van Der Hoek, W. (2014). Structures for epistemic logic. In *Johan van Benthem on logic and information dynamics*, pages 339–380. Springer.

Birch, S. A. and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5):382–386. PMID: 17576275.

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

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

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

Board, O. (2004). Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49 – 80.

Bolander, T. (2018). Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In van Ditmarsch, H. and Sandu, G., editors, *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 207–236. Springer.

Bonnay, D. and Égré, P. (2009). Inexact knowledge with introspection. *Journal of Philosophical Logic*, 38(2):179–227.

Bonnay, D. and Égré, P. (2011). *Knowing One's Limits: An Analysis in Centered Dynamic Epistemic Logic*, pages 103–126. Springer.

Bouadjenek, M. R., Pacitti, E., Servajean, M., Masseglia, F., and Abbadi, A. E. (2018). A distributed collaborative filtering algorithm using multiple data sources. *arXiv preprint arXiv:1807.05853*.

Bradford, E. E., Jentzsch, I., and Gomez, J.-C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. *Cognition*, 138:21 – 34.

Braine, M. D., O'Brien, D. P., and Braine, M. (1998). Mental logic and irrationality: We can put a man on the moon, so why can't we solve those logical reasoning problems? In *Mental logic*, pages 31–52. Psychology Press.

Braüner, T. (2014). Hybrid-logical reasoning in the smarties and Sally-Anne tasks. *Journal of Logic, Language and Information*, 23(4):415–439.

Braüner, T. (2015). Hybrid-logical reasoning in the smarties and Sally-Anne tasks: What goes wrong when incorrect responses are given? In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, Pasadena, California, USA*, pages 273–278. Cognitive Science Society.

Braüner, T. (2017). Hybrid logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Braüner, T., Blackburn, P., and Polyanskaya, I. (2016). Second-order false-belief tasks: Analysis and formalization. In *23rd International Workshop, WoLLIC 2016*, pages 125–144. Springer.

Burgess, J. P. (1982). Axioms for tense logic. I. "since" and "until". *Notre Dame Journal of Formal Logic*, 23(4):367–374.

Butler, J. (2013). *Rethinking Introspection: A Pluralist Approach to the First-Person Perspective.* Palgrave MacMillan.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31(1):61 – 83.

Carruthers, P. and Smith, P. K. (1996). *Theories of Theories of Mind.* CUP.

Charrier, T., Herzig, A., Lorini, E., Maffre, F., and Schwarzentruber, F. (2016). Building epistemic logic from observations and public announcements. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'16, pages 268–277. AAAI Press.

Cheng, P. W. and Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17(4):391 – 416.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* The MIT Press, 50 edition.

Clark, H. H. and Marshall, C. R. (1981). Definite knowledge and mutual knowledge. In Joshi, A. K., Webber, B. L., and Sag, I. A., editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge, UK: Cambridge University Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4(3):317–331.

Cohen, L. J. (1994). A reply to Stein. *Synthese*, 99(2):173–176.

Colyvan, M. (2013). Idealisations in normative models. *Synthese*, 190(8):1337–1350.

Coppee, H. (1860). *Elements of Logic; Designed as a manual of instruction.*

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3):187 – 276.

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, 163:163–228.

Counihan, M. E. et al. (2008). *Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning.* Institute for Logic, Language and Computation.

Cowan, N. (1999). An embedded-processes model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, 20:506.

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.

Cozic, M. (2006). Impossible states at work: Logical omniscience and rational choice. *Contributions to Economic Analysis*, 280:47–68.

Daniels, N. (2020). Reflective equilibrium. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition.

Danto, A. C. (1967). On knowing that we know. In Stroll (1967), pages 32–53.

Davidson, D. (1967). Truth and meaning. *Synthese*, 17(1):304–323.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47:5–20.

Demey, L., Kooi, B., and Sack, J. (2019). Logic and Probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Dennett, D. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1981). True believers: The intentional strategy and why it works. In Heath, A. F., editor, *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, pages 150–167. Clarendon Press.

Dieussaert, K., Schaeken, W., Schroyens, W., and D'Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking & Reasoning*, 6(2):125–160.

Dissing, L. and Bolander, T. (2020). Implementing theory of mind on a robot using dynamic epistemic logic. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,(IJCAI 2020)*, pages 1615–1621.

Dretske, F. (2004). Externalism and modest contextualism. *Erkenntnis (1975-)*, 61(2/3):173–186.

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

Dunin-Kęplicz, B. and Verbrugge, R. (2006). Awareness as a vital ingredient of teamwork. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '06, pages 1017–1024, New York, NY, USA. ACM.

Dunin-Kęplicz, B. and Verbrugge, R. (2012). A logical view on teamwork. In *Games, Actions and Social Software*, pages 184–212. Springer.

D'Agostino, M. (2010). Tractable depth-bounded logics and the problem of logical omniscience. *Probability, uncertainty and rationality*, pages 245–275.

D'Agostino, M. and Floridi, L. (2009). The enduring scandal of deduction. *Synthese*, 167(2):271–315.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds and Markets: Reasoning about a Highly Connected World.* Cambridge University Press, New York.

Égré, P. (2020). Logical omniscience. *The Wiley Blackwell Companion to Semantics*, pages 1–25.

Elgot-Drapkin, J., Kraus, S., Miller, M., Nirkhe, M., and Perlis, D. (1999). Active logics: A unified formal approach to episodic reasoning.

Elqayam, S. (2018). The new paradigm in psychology of reasoning. In Ball, L. and Thompson, V., editors, *The Routledge International Handbook of Thinking and Reasoning*, pages 130–50. Routledge.

Epley, N., Keysar, B., Boven, L. V., and Gilovich, T. D. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, 87 3:327–39.

Evans, J. S., Legrenzi, P., and Girotto, V. (1999). The influence of linguistic form on reasoning: The case of matching bias. *The Quarterly Journal of Experimental Psychology Section A*, 52(1):185–216.

Evans, J. S. B., Over, D. E., and Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition*, 49(1-2):165–187.

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468.

Evans, J. S. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, 4(1):45–110.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454 – 459.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59(1):255–278. PMID: 18154502.

Evans, J. S. B. T. (2018). Dual process theories. In Ball, L. and Thompson, V., editors, *The Routledge International Handbook of Thinking and Reasoning*, pages 151–64. Routledge.

Evans, J. S. B. T., Barston, J. L., and Pollard, P. T. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11:295–306.

Evans, J. S. B. T. and Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4):382–389.

Evans, J. S. B. T. and Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology*, 64(3):391–397.

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

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

Fagin, R., Halpern, J. Y., and Megiddo, N. (1990). A logic for reasoning about probabilities. *Information and Computation*, 87(1):78 – 128. Special Issue: Selections from 1988 IEEE Symposium on Logic in Computer Science.

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

Fagin, R., Halpern, J. Y., and Vardi, M. Y. (1992). What can machines know? on the properties of knowledge in distributed systems. *J. ACM*, 39(2):328–376.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633.

Feldman, J. (2003). A catalog of boolean concepts. *J. Math. Psychol.*, 47(1):75–89.

Fernández-Fernández, C. and Velázquez-Quesada, F. R. (2019). Awareness of and awareness that: their combination and dynamics. *Logic Journal of the IGPL*.

Fervari, R. and Velázquez-Quesada, F. R. (2019). Introspection as an action in relational models. *Journal of Logical and Algebraic Methods in Programming*, 108:1–23.

Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442.

Floridi, L. (2005). Is information meaningful data? *Philosophy and Phenomenological Research*, 70(2):351–70.

Fodor, J. A. (1981). Three cheers for propositional attitudes. In Fodor, J. A., editor, *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.

Frege, G. (1884). *Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner.

Frigg, R. (2010). Models and fiction. *Synthese*, 172(2):251–268.

Gärdenfors, P. (1988). *Knowledge in Flux. Modelling the Dynamics of Epistemic States*. MIT Press.

Gasquet, O., Goranko, V., and Schwarzentruber, F. (2016). Big brother logic: visual-epistemic reasoning in stationary multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 30(5):793–825.

Gaudou, B., Herzig, A., Longin, D., and Lorini, E. (2015). On modal logics of group belief. In *The Cognitive Foundations of Group Attitudes and Social Interaction*, pages 75–106. Springer.

Geil, D. M. M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, 4(3):231–248.

Gerbrandy, J. D. (1999). *Bisimulations on planet Kripke*. Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Ghosh, S., Meijering, B., and Verbrugge, R. (2014). Strategic reasoning: Building cognitive models from logical formulas. *Journal of Logic, Language and Information*, 23(1):1–29.

Ghosh, S. and Verbrugge, R. (2018). Studying strategies and types of players: Experiments, logics and cognitive models. *Synthese*, 195(10):4265–4307.

Gierasimczuk, N., van der Maas, H. L. J., and Raijmakers, M. E. J. (2013). An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, 22(3):297–314.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, 2(1):83–115.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, 143(1):101–116.

Goldman, A. I. (1978). Epistemics: The regulative theory of cognition. *Journal of Philosophy*, 75(10):509–523.

Goodman, N. (1955). *Fact, Fiction and Forecast.* Harvard University Press.

Goodwin, G. P. and Johnson-Laird, P. N. (2013). The acquisition of boolean concepts. *Trends in cognitive sciences*, 17(3):128–133.

Gopnik, A. and Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 59(1):26–37.

Grandy, R. (1973). Reference, meaning, and belief. *Journal of Philosophy*, 70(14):439–452.

Griggs, R. A. (1995). The effects of rule clarification, decision justification, and selection instruction on Wason's abstract selection task.

Griggs, R. A. and Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 73(3):407–420.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170.

Gruber, H. (1990). The cooperative synthesis of disparate points of view. *The legacy of Solomon Asch: Essays in cognition and social psychology*, 1:143–158.

Hájek, A. (2001). Probability, logic, and probability logic. *The Blackwell guide to philosophical logic*, pages 362–384.

Halpern, J. Y. (1996). Should knowledge entail belief? *Journal of Philosophical Logic*, 25(5):483–494.

Halpern, J. Y. (2017). *Reasoning about uncertainty.* MIT press.

Halpern, J. Y. and Pucella, R. (2011). Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence*, 175(1):220 – 235. John McCarthy's Legacy.

Harel, D., Kozen, D., and Tiuryn, J. (2000). *Dynamic Logic.* MIT Press, Cambridge, USA.

Harman, G. (1991). Justification, truth, goals, and pragmatism: Comments on Stich's fragmentation of reason. *Philosophy and Phenomenological Research*, 51(1):195–199.

Hawke, P., Özgün, A., and Berto, F. (2019). The fundamental problem of logical omniscience. *Journal of Philosophical Logic*, pages 1–40.

Hedden, T. and Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36.

Hendricks, V. F. and Symons, J. E. (2006). Where's the bridge? Epistemology and Epistemic Logic. *Philosophical Studies*, 128:137–167.

Henle, M. (1962). On the relation between logic and thinking. *Psychological review*, 69(4):366.

Hertwig, R. and Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12(4):275–305.

Herzig, A., Lorini, E., and Maffre, F. (2018). Possible worlds semantics based on observation and communication. In van Ditmarsch, H. and Sandu, G., editors, *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 339–362. Springer.

Hintikka, J. (1962). *Knowledge and Belief.* Ithaca: Cornell University Press.

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

Hocutt, M. O. (1972). Is epistemic logic possible? *Notre Dame J. Formal Logic*, 13(4):433–453.

Hoek, D. (2020). Minimal rationality and the web of questions. In Kindermann, D., van Elswyk, P., and Egan, A., editors, *Unstructured Content.* Oxford University Press.

Hollebrandse, B., van Hout, A., and Hendriks, P. (2014). Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. *Synthese*, 191(3):321–333.

Holyoak, K. J., Cheng, P. W., Newstead, S., and Evans, S. J. (1995). Pragmatic reasoning about human voluntary action: Evidence from Wason's selection task. *Perspectives on thinking and reasoning: Essays in honour of Peter Wason*, pages 67–89.

Icard, T. F. (2014). *The algorithmic mind: A study of inference in action.* PhD thesis, Stanford University.

Inglis, M. and Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43(4):358–390.

Isaac, A. M., Szymanik, J., and Verbrugge, R. (2014). Logic and complexity in cognitive science. *Johan van Benthem on logic and information dynamics*, pages 787–824.

Jago, M. (2006). Hintikka and Cresswell on logical omniscience. *Logic and Logical Philosophy*, 15(4):325–354.

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

Jago, M. (2013). The problem of rational knowledge. *Erkenntnis*, pages 1–18.

Jago, M. (2014). *The Impossible: An Essay on Hyperintensionality.* Oxford University Press.

Jarrold, C. and Towse, J. (2006). Individual differences in working memory. *Neuroscience*, 139(1):39 – 50.

Johansson, P., Hall, L., Sikström, S., Tärning, B., and Lind, A. J. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition*, 15:673–692.

Johnson-Laird, P. (1983). *Mental Models.* Cambridge University Press.

Johnson-Laird, P. N. (2013). *Human and machine thinking.* Psychology Press.

Johnson-Laird, P. N. and Byrne, R. M. (1991). *Deduction.* Lawrence Erlbaum Associates, Inc.

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

Johnson-Laird, P. N., Legrenzi, P., and Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, 63(3):395–400.

Just, M. A. and Carpenter, P. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99 1:122–49.

Kahneman, D. (1973). *Attention and Effort.* Prentice-Hall.

Kahneman, D. (1981). Who shall be the arbiter of our intuitions? *Behavioral and Brain Sciences*, 4(3):339–340.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux, New York.

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

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291.

Kamp, H. (1968). *Tense Logic and the Theory of Linear Order*. PhD thesis, University of California.

Kasbergen, O. (2017). Abstractions and idealizations in epistemic logic. Master's thesis, University of Amsterdam, the Netherlands.

Katz, D. and Allport, F. H. (1931). *Student Attitudes*. Syracuse, N.Y.: Craftsman.

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1):25–41.

Kinderman, P., Dunbar, R., and Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2):191–204.

Klein, D. (2015). *Social interaction: A formal exploration*. PhD thesis, Tilburg University.

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

Kurtonina, N. and De Rijke, M. (1997). Bisimulations for temporal logic. *Journal of Logic, Language and Information*, 6(4):403–425.

Lane, K. A., Banaji, M. R., Nosek, B. A., and Greenwald, A. G. (2007). Understanding and using the implicit association test: IV. *Implicit measures of attitudes*, pages 59–102.

Laughlin, P. R. and Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, 22(3):177 – 189.

Lehrer, K. (2000). *Theory of Knowledge*. Westview Press.

Lehrer, K. and Paxson, T. (1969). Knowledge: Undefeated justified true belief. *Journal of Philosophy*, 66(8):225–237.

Leitgeb, H. (2008). Introduction to the special issue. *Studia Logica*, 88(1):1–2.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press.

Lemmon, E. J. (1967). If I know, do I know that I know? In Stroll (1967), pages 54–83.

Lemmon, E. J. and Henderson, G. P. (1959). Is there only one correct system of modal logic? *Aristotelian Society Supplementary Volume*, 33(1):23–56.

Lenzen, W. (1978). Recent work in epistemic logic.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the Fourth AAAI Conference on Artificial Intelligence*, AAAI'84, pages 198–202. AAAI Press.

Levi, I. (1983). Who commits the base rate fallacy? *Behavioral and Brain Sciences*, 6(3):502–506.

Lin, S., Keysar, B., and Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46(3):551 – 556.

Litman, L. and Reber, A. S. (2005). Implicit cognition and thought. In Holyoak, K. and Morrison, B., editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 431–453. Cambridge University Press.

Liu, F. (2008). *Changing for the Better: Preference Dynamics and Agent Diversity.* PhD thesis, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands.

Liu, F. (2009). Diversity of agents and their interaction. *Journal of Logic, Language and Information*, 18(1):23–53.

Liu, F. (2011). *Reasoning About Preference Dynamics.* Springer Verlag.

Lorini, E. (2020). Rethinking epistemic logic with belief bases. *Artificial Intelligence*, 282:103233.

Macnamara, J. (1990). A border dispute: The place of logic in psychology. *Philosophy of Science*, 57(2):347–349.

Manktelow, K. I. and Evans, J. S. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology*, 70(4):477–488.

Marcus, S. L. and Rips, L. J. (1979). Conditional reasoning. *Journal of Verbal Learning and Verbal Behavior*, 18(2):199 – 223.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., New York, NY, USA.

Meijering, B. (2014). *Reasoning about self and others.* PhD thesis, University of Groningen.

Meijering, B., Taatgen, N. A., van Rijn, H., and Verbrugge, R. (2014). Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies*, 15(3):455–477.

Meijering, B., Van Maanen, L., Van Rijn, H., and Verbrugge, R. (2010). The facilitative effect of context on second-order social reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Meijering, B., van Rijn, H., Taatgen, N., and Verbrugge, R. (2011). I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 2486–2491.

Meijering, B., van Rijn, H., Taatgen, N. A., and Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PLOS ONE*, 7(9):1–8.

Meijering, B., Van Rijn, H., Taatgen, N. A., and Verbrugge, R. (2013). Reasoning about diamonds, gravity and mental states: The cognitive costs of theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Messer, W. S. and Griggs, R. A. (1993). Another look at Linda. *Bulletin of the Psychonomic Society*, 31(3):193–196.

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

Moreno, A. (1998). Avoiding logical omniscience and perfect reasoning: a survey. *AI Communications*, 11(2):101–122.

Nguyen, H. N. and Rakib, A. (2019). A probabilistic logic for resource-bounded multi-agent systems. In *IJCAI*, pages 521–527.

Nichols, S. and Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford University Press.

Nisbett, R. E. and Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of personality and social psychology*, 35(9):613.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review; Psychological Review*, 84(3):231.

Nolan, D. (1997). Impossible worlds: A modest approach. *Notre Dame J. Formal Logic*, 38(4):535–572.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory.* The MIT Press, Cambridge, Massachusetts.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11(6):988–1010.

Parikh, R. (1987). Knowledge and the problem of logical omniscience. In *Proceedings of the Second International Symposium on Methodologies for Intelligent Systems*, pages 432–439, Amsterdam, The Netherlands. North-Holland Publishing Co.

Parikh, R. (2007). Logical omniscience in the many agent case. Technical report, The City University of New York.

Pelletier, F. J., Elio, R., and Hanson, P. (2008). Is logic all in our heads? From naturalism to psychologism. *Studia Logica*, 88(1):3–66.

Perner, J. and Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39:437–471.

Piaget, J. (1953). *Logic and Psychology.* New York: Basic Books.

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

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Priest, G. (2006). *In Contradiction.* Oxford University Press UK.

Proietti, C. and Olsson, E. J. (2014). A DDL approach to pluralistic ignorance and collective belief. *Journal of Philosophical Logic*, 43(2):499–515.

Punčochář, V. and Sedlár, I. (2017). Substructural logics for pooling information. In Baltag, A., Seligman, J., and Yamada, T., editors, *Logic, Rationality, and Interaction*, pages 407–421, Berlin, Heidelberg. Springer Berlin Heidelberg.

Quine, W. V. O. (1960). *Word & Object.* MIT Press.

Quine, W. V. O. (1969). Epistemology naturalized. In *Ontological Relativity and Other Essays.* New York: Columbia University Press.

Rahman, S., Symons, J., Gabbay, D. M., and van Bendegem, J. P. (2004). *Logic, epistemology, and the unity of science.* Springer.

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

Rantala, V. (1982b). Quantified modal logic: Non-normal worlds and propositional attitudes. *Studia Logica*, 41:41–65.

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

Rawls, J. (2009). *A theory of justice*. Harvard university press.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6(6):855 – 863.

Reber, A. S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. *Oxford University Press*.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132.

Rendsvig, R. and Symons, J. (2019). Epistemic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Rendsvig, R. K. (2014). Pluralistic ignorance in the bystander effect: Informational dynamics of unresponsive witnesses in situations calling for intervention. *Synthese*, 191(11):2471–2498.

Renne, B., Sack, J., and Yap, A. (2016). Logics of temporal-epistemic actions. *Synthese*, 193(3):813–849.

Rijmen, F. and De Boeck, P. (2001). Propositional reasoning: The differential contribution of "rules" to the difficulty of complex reasoning problems. *Memory & Cognition*, 29(1):165–175.

Rips, L. and Conrad, F. (1983). Individual differences in deduction. *Cognition and Brain Theory*, 6:259–285.

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

Roberts, M. J. and Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *The Quarterly Journal of Experimental Psychology Section A*, 54(4):1031–1048. PMID: 11765731.

Roelofsen, F. (2007). Distributed knowledge. *Journal of Applied Non-Classical Logics*, 17(2):255–273.

Rott, H. (1989). Conditionals and theory change: Revisions, expansions, and additions. *Synthese*, 81(1):91–113.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, 61(2-3):469–493.

Rubinstein, A. (1989). The Electronic Mail Game: Strategic Behavior under "Almost Common Knowledge". *American Economic Review*, 79(3):385–391.

Sack, J. (2008). Temporal languages for epistemic programs. *Journal of Logic, Language and Information*, 17(2):183–216.

Schacter, D. L. and Tulving, E. (1994). *Memory Systems 1994*. MIT Press.

Schunn, C. D. and Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory & Cognition*, 24(3):271–284.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91(4):531–553.

Seligman, J., Liu, F., and Girard, P. (2013). Facebook and the epistemic logic of friendship. In Schipper, B. C., editor, *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2013), Chennai, India, January 7-9, 2013*.

Setoh, P., Scott, R. M., and Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences*, 113(47):13360–13365.

Sillari, G. (2008). Quantified logic of awareness and impossible possible worlds. *The Review of Symbolic Logic*, 1(4):514–529.

Sim, K. M. (1997). Epistemic logic and logical omniscience: A survey. *International Journal of Intelligent Systems*, 12(1):57–81.

Skipper, M. and Bjerring, J. C. (2020). Bayesianism for the average Joe. Preprint.

Smets, S. and Solaki, A. (2018). The effort of reasoning: Modelling the inference steps of boundedly rational agents. In *International Workshop on Logic, Language, Information, and Computation*, pages 307–324. Springer.

Smets, S. and Velázquez-Quesada, F. R. (2017). How to make friends: A logical approach to social group creation. In *International Workshop on Logic, Rationality and Interaction*, pages 377–390. Springer.

Sober, E. (1981). The evolution of rationality. *Synthese*, 46(January):95–120.

Solaki, A. (2017). Steps out of logical omniscience. Master's thesis, University of Amsterdam.

Solaki, A. (2018). Rule-based reasoners in epistemic logic. In *European Summer School in Logic, Language and Information*, pages 144–156. Springer.

Solaki, A. (2019). A dynamic epistemic logic for resource-bounded agents. In Sedlar, I. and Blicha, M., editors, *The Logica Yearbook 2018*, pages 229–244. College Publications.

Solaki, A. (2020). Bounded multi-agent reasoning: Actualizing distributed knowledge. In *International Workshop on Dynamic Logic*, pages 239–258. Springer.

Solaki, A. (2021a). Actualizing distributed knowledge in bounded groups. Submitted manuscript.

Solaki, A. (2021b). Bounded multi-agent reasoning: Inference, introspection, attribution. Submitted manuscript.

Solaki, A. (2021c). Where is Epistemic Logic in the Rationality Debate? Submitted manuscript.

Solaki, A., Berto, F., and Smets, S. (2019). The logic of fast and slow thinking. *Erkenntnis*, pages 1–30.

Solaki, A. and Smets, S. (2021). The effort of reasoning: Modelling the inference steps of boundedly rational agents. *Journal of Logic, Language and Information.* (forthcoming).

Solaki, A. and Velázquez-Quesada, F. R. (2019). Towards a logical formalisation of theory of mind: A study on false belief tasks. In *International Workshop on Logic, Rationality and Interaction*, pages 297–312. Springer.

Solaki, A. and Velázquez-Quesada, F. R. (2020). What do you believe your friends believe? Towards realistic belief attributions in multi-agent systems. In *NETREASON workshop, 24th European Conference on Artificial Intelligence.* (Extended abstract).

Solaki, A. and Velázquez-Quesada, F. R. (2021). A logical formalisation of false belief tasks. Submitted manuscript.

Solomon, M. (1994). Stich's The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation. *Informal Logic*, 16(2).

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

Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese*, 89(3):425–440.

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

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. *The Oxford Handbook of Thinking and Reasoning.*

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

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science.* Clarendon Press.

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

Stich, S. P. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation.* The MIT Press.

Stich, S. P. and Nisbett, R. E. (1980). Justification and the psychology of human reasoning. *Philosophy of Science*, 47(2):188–202.

Stiller, J. and Dunbar, R. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93 – 104.

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

Szymanik, J., Meijering, B., and Verbrugge, R. (2013). Using intrinsic complexity of turn-taking games to predict participants' reaction times. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Thagard, P. (1982). From the descriptive to the normative in psychology and logic. *Philosophy of Science*, 49(1):24–42.

Thagard, P. and Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science*, 50(2):250–267.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

Thaler, R. H. (2012). *The winner's curse: Paradoxes and anomalies of economic life.* Simon and Schuster.

Thijsse, E. (1993). On total awareness logics. In *Diamonds and Defaults*, pages 309–347. Springer.

Trognon, A., Batt, M., and Laux, J. (2011). Why is dialogical solving of a logical problem more effective than individual solving?: A formal and experimental study of an abstract version of Wason's task. *Language and Dialogue*, 1(1):44–78.

Tversky, A. and Kahneman, D. (1975). *Judgment under Uncertainty: Heuristics and Biases*, pages 141–162. Springer Netherlands, Dordrecht.

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

Tversky, A. and Kahneman, D. (1985). *The Framing of Decisions and the Psychology of Choice*, pages 107–129. Springer Berlin Heidelberg, Berlin, Heidelberg.

van Benthem, J. (2004). A mini-guide to logic in action. *Tech Report ILLC PP*, (02).

van Benthem, J. (2006). Epistemic logic and epistemology: The state of their affairs. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 128(1):49–76.

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

van Benthem, J. (2008a). Logic and reasoning: Do the facts matter? *Studia Logica: An International Journal for Symbolic Logic*, 88(1):67–84.

van Benthem, J. (2008b). Merging observation and access in dynamic epistemic logic. *Studies in Logic*, 1:1–16.

van Benthem, J. (2008c). '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. (2017). *Against All Odds: When Logic Meets Probability*, pages 239–253. Springer International Publishing, Cham.

van Benthem, J., Hodges, H., and Hodges, W. (2007). Topos: Logic and cognition-introduction.

van Benthem, J. and Liu, F. (2007). Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182.

van Benthem, J., Liu, F., and Smets, S. (2020). Logico-computational aspects of rationality. *"Handbook of Rationality", M. Knauff & W. Spohn, eds.*

van Benthem, J., Martinez, M., Israel, D., and Perry, J. (2008). The stories of logic and information. *Handbook of the Philosophy of Information, Elsevier Science Publishers, Amsterdam*, pages 217–280.

van Benthem, J., van Eijck, J., Gattinger, M., and Su, K. (2018). Symbolic model checking for dynamic epistemic logic - S5 and beyond. *Journal of Logic and Computation*, 28(2):367–402.

van Benthem, J., van Eijck, J., and Kooi, B. (2006). Logics of communication and change. *Information and Computation*, 204(11):1620 – 1662.

Van De Pol, I., Van Rooij, I., and Szymanik, J. (2018). Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information*, 27(3):255–294.

van der Hoek, W., Troquard, N., and Wooldridge, M. J. (2011). Knowledge and control. In Sonenberg, L., Stone, P., Tumer, K., and Yolum, P., editors, *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011, Volume 1-3*, pages 719–726. IFAAMAS.

van Ditmarsch, H., Halpern, J., van der Hoek, W., and Kooi, B. (2015). *Handbook of Epistemic Logic*. College Publications.

van Ditmarsch, H. and Kooi, B. (2008). Semantic results for ontic and epistemic change. In Bonanno, G., van der Hoek, W., and Wooldridge, M., editors, *Logic and the Foundations of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 87–117. Amsterdam University Press, Amsterdam, The Netherlands.

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

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

van Lambalgen, M. and Hamm, F. (2008). *The proper treatment of events*, volume 6. John Wiley & Sons.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, 32(6):939–984.

Velázquez-Quesada, F. R. (2009). Inference and update. *Synthese*, 169(2):283–300.

Velázquez-Quesada, F. R. (2011). *Small Steps in Dynamics of Information*. PhD thesis, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands.

Velázquez-Quesada, F. R. (2013). Explicit and implicit knowledge in neighbourhood models. In *International Workshop on Logic, Rationality and Interaction*, pages 239–252. Springer.

Velázquez-Quesada, F. R. (2014). Dynamic epistemic logic for implicit and explicit beliefs. *Journal of Logic, Language and Information*, 23(2):107–140.

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

Verbrugge, R. and Mol, L. (2008). Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17(4):489–511.

Vidal, J. M. (2004). A protocol for a distributed recommender system. In *Trusting agents for trusting electronic societies*, pages 200–217. Springer.

Vigo, R. (2006). A note on the complexity of boolean concepts. *Journal of Mathematical Psychology*, 50(5):501–510.

von Wright, G. H. (1951). *An Essay in Modal Logic.* Amsterdam: North-Holland Pub. Co.

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

Wason, P. and Johnson-Laird, P. (1972). *Psychology of reasoning: structure and content.* Harvard paperback. Harvard University Press.

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

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281. PMID: 5683766.

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

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for research in mathematics education*, pages 431–459.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, 104(12):639–659.

Wellman, H. M. (1991). From desires to beliefs: Acquisition of a theory of mind. In *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, pages 19–38. Basil Blackwell, Cambridge, MA, US.

Williamson, T. (1992). Inexact knowledge. *Mind*, 101(402):217–242.

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

Williamson, T. (2001). Some philosophical aspects of reasoning about knowledge. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, pages 97–97. Morgan Kaufmann Publishers Inc.

Williamson, T. (2020). *Suppose and tell: The semantics and heuristics of conditionals*. Oxford University Press.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103 – 128.

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

Yap, A. (2011). Dynamic epistemic logic and temporal modality. In Girard, P., Roy, O., and Marion, M., editors, *Dynamic Formal Epistemology*, pages 33–50. Springer.

Yap, A. (2014). Idealization, epistemic logic, and epistemology. *Synthese*, 191(14):3351–3366.

Zhai, F., Szymanik, J., and Titov, I. (2015). Toward probabilistic natural logic for syllogistic reasoning. In *Proceedings of the 20th Amsterdam Colloquium*, pages 468–477.

Zhao, B., van de Pol, I., Raijmakers, M., and Szymanik, J. (2018). Predicting cognitive difficulty of the deductive mastermind game with dynamic epistemic logic models. In Kalish, C., Rau, M., and Zhu, J., editors, *CogSci 2018 - 40th Annual Cognitive Science Society Meeting [Proceedings]*, pages 2789–2794. Cognitive Science Society.

Ågotnes, T. and Alechina, N. (2006). The Dynamics of Syntactic Knowledge. *Journal of Logic and Computation*, 17(1):83–116.

Ågotnes, T. and Walicki, M. (2004). Syntactic knowledge: A logic of reasoning, communication and cooperation. In *Proceedings of the Second European Workshop on Multi-Agent Systems (EUMAS), Barcelona, Spain.*

Ågotnes, T. and Wáng, Y. N. (2017). Resolving distributed knowledge. *Artificial Intelligence*, 252:1 – 21.

Abstract

Logical Models for Bounded Reasoners

This dissertation aims at the logical modelling of aspects of human reasoning, informed by facts on the bounds of human cognition. We break down this challenge into three parts: Part I, explaining why the design of such logical systems is a worthwhile project; Parts II and III, providing logical frameworks to this end, concerning, respectively, *single-agent* and *multi-agent* reasoning.

In Part I, we discuss the place of logical systems for knowledge and belief in the *Rationality Debate*, i.e. the debate on whether humans are rational or not. We argue for the need to revise the standard epistemic/doxastic logics (**S5**/**KD45**, respectively) in order to provide formal counterparts of an alternative picture of rationality – one wherein empirical facts have a key role (Chapter 2).

In Part II, we design *resource-sensitive* logical models that encode explicitly the deductive reasoning of a bounded agent and the variety of processes underlying it. This is achieved through the introduction of a dynamic, impossible-worlds semantics, with quantitative components capturing the agent's cognitive capacity and the cognitive costs of deductive inference rules with respect to certain resources, such as memory and time (Chapter 3). We then show that this type of semantics can be combined with plausibility models, which allow for (i) the study of more nuanced notions of knowledge and belief from the resource-sensitive perspective, and (ii) the study of the interplay between inference and interaction (Chapter 4). We proceed with the demonstration of another contribution of this type of semantics; we show it can be instrumental in modelling the logical aspects of *System 1* ("fast") and *System 2* ("slow") cognitive processes, as per dual process theories of reasoning (Chapter 5).

In Part III, we move from single- to multi-agent frameworks. This unfolds in three directions: (a) the formation of beliefs about others (e.g. due to observation, memory, and communication), (b) the manipulation of beliefs (e.g. via acts of reasoning about oneself and others), and (c) the effect of the above on group reasoning. Point (a) is addressed through the design of temporal models keeping track of agents' visibility and communication; the framework is applied to the formalization of paradigmatic tasks testing people's *Theory of Mind*, the so-called *False Belief Tasks* (Chapter 6). Point (b) is addressed through the design of special action models, which are compatible with our resource-sensitive semantics and able to represent actions of deduction, introspection, and attribution, that, when cognitively affordable, can refine the zero- and higher-order beliefs of agents (Chapter 7). Point (c) is addressed by first looking into idealizations of group epistemic notions, with an emphasis on *distributed knowledge*. Inspired by experiments on group reasoning, we then identify two dimensions of actualizing distributed knowledge under bounded resources, namely communic-ation and inference. Using the toolbox introduced earlier, we build a dynamic framework with effortful actions accounting for both (Chapter 8).

We finally discuss directions for future work, touching upon the study of probabilistic reasoning and social networks, and we reflect on the contribution of the thesis as a whole (Chapter 9).

Samenvatting

Logische Modellen voor Begrensde Denkers

Dit proefschrift richt zich op de logische modellering van aspecten van het menselijk redeneren, gebaseerd op feiten met betrekking tot de grenzen van de menselijke cognitie. We splitsen deze uitdaging op in drie delen: Deel I legt uit waarom het creëren van dergelijke logische systemen een waardevol project is; Delen II en III verschaffen hiervoor logische kaders, respectievelijk betreffende *single-agent* en *multi-agent* redenering.

In Deel I, bespreken we de plaats van logische systemen voor kennis en geloof in het *rationaliteitsdebat*, dat wil zeggen het debat over de vraag of mensen rationeel zijn of niet. We pleiten voor de noodzaak om de standaard epistemische/doxastische logica (respectievelijk **S5**/**KD45**) te herzien om formele tegenhangers te bieden van een alternatief beeld van rationaliteit - één waarin empirische feiten bepalend zijn (Hoofdstuk 2).

In Deel II, ontwerpen we logische modellen die rekening houden met de cognitieve bronnen van een begrensde agent (i.e., die *'bron-gevoelig'* is) en die expliciet de deductieve redenering van een begrensde agent en de verscheidenheid van onderliggende processen coderen. Dit wordt bereikt door de introductie van een dynamische semantiek met onmogelijke werelden, waarbij kwantitatieve componenten de cognitieve capaciteit van de agent en de cognitieve kosten van deductieve inferentieregels met betrekking tot bepaalde bronnen, zoals geheugen en tijd, vastleggen (Hoofdstuk 3). We laten vervolgens zien dat dit type semantiek kan worden gecombineerd met plausibiliteitsmodellen, die het mogelijk maken om (i) meer genuanceerde noties van kennis en geloof vanuit het bron-gevoelige perspectief te bestuderen, en (ii) de wisselwerking tussen inferentie en communicatie te bestuderen (Hoofdstuk 4). We tonen verder een andere bijdrage van dit soort semantiek aan; we laten zien dat het instrumenteel kan zijn bij het modelleren van de logische aspecten van *Systeem 1* ("snelle") en *Systeem 2* ("langzame") cognitieve processen, in lijn met de duale procestheorieën van redeneren (Hoofdstuk 5).

In Deel III, gaan we van single-agent- naar multi-agent kaders. Dit ontvouwt zich in drie richtingen: (a) de vorming van overtuigingen over anderen (bijv. door observatie, geheugen en communicatie), (b) de manipulatie van overtuigingen (bijv. via redenering over zichzelf en anderen) en (c) het effect van het bovenstaande op het groepsredeneren. Punt (a) wordt aangepakt door modellen die de zichtbaarheid en communicatie van agenten in de tijd bijhouden; het raamwerk wordt toegepast op de formalisering van paradigmatische taken die de '*Theory of Mind*' van mensen testen, de zogenaamde '*False Belief Tasks*' (Hoofdstuk 6). Punt (b) wordt aangepakt door het ontwerp van speciale actiemodellen, die compatibel zijn met onze bron-gevoelige semantiek en in staat zijn om acties van deductie, introspectie en attributie weer te geven, die, indien cognitief haalbaar, de nul- en hogere-orde overtuigingen van agenten kunnen verfijnen (Hoofdstuk 7). Punt (c) wordt aangepakt door eerst te kijken naar idealisaties van groep epistemologische noties, met de nadruk op *gedistribueerde kennis.* Geïnspireerd door experimenten met betrekking tot groepsredenering, identificeren we twee dimensies van het actualiseren van gedistribueerde kennis met begrensde bronnen, namelijk communicatie en inferentie. Met behulp van het eerder geïntroduceerde formele gereedschap construeren we een dynamisch raamwerk met inspanningsvolle acties die voor zowel communicatie als inferentie zorgen (Hoofdstuk 8).

Als laatste bespreken we mogelijkheden voor toekomstig werk, waarbij we de studie van probabilistisch redeneren en sociale netwerken aanraken. We reflecteren ook op de bijdrage van het proefschrift als geheel (Hoofdstuk 9).

Titles in the ILLC Dissertation Series:

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility

ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms

ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation

ILLC DS-2021-02: **Elbert J.Booij**
The Things Before Us: On What it Is to Be an Object

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices