Proceedings of the

Workshop on Logics for Resource-Bounded Agents

organised as part of the 18th European Summer School on Logic, Language and Information (ESSLLI), August 2006, Malaga, Spain

Thomas Ågotnes and Natasha Alechina (eds.)

Foreword

Logics of knowledge and belief, as well as other attitudes such as desire or intention, have been extensively studied. However, most of the treatments of knowledge and belief make strong and idealised assumptions about the reasoners. For example, traditional epistemic logic says that agents know all logical consequences of their knowledge. Similarly, logics of action and strategic interaction are usually based on game theoretic models which assume perfect rationality. Models based on such assumptions can be used to describe ideal agents without bounds on resources such as time, memory, etc, but they fail to accurately describe non-ideal agents which are computationally bounded.

The Workshop on Logics for Resource-Bounded Agents was held as a part of the 18th European Summer School in Logic, Language and Information (ESSLLI, Malaga, 2006), in order to provide a forum for advanced PhD students and researchers to present and discuss possible solutions to the problem of formally capturing the properties of knowledge, belief, action, etc., of non-idealised resource-bounded agents with colleagues and researchers who work in logic, computer science and other areas represented at ESSLLI.

The workshop started with an invited talk by Rohit Parikh on *Belief, Knowledge, Sentences and Propositions*.

10 contributed papers were accepted for presentation at the workshop. Active logics, a logical framework for modeling resource bounded agents, receive attention from several of the authors. Mikael Asker and Jacek Malec discuss, in the paper *Improving active consequence*, notions of semantic consequence in active logics. In *A modal logic of time-bounded reasoning agents*, Ayelet Butman and Sarit Kraus present a modal active logic, describing how the beliefs of non-ideal reasoners can evolve over time. Slawomir Nowaczyk also uses active logic; in *Partial planning for situated agents based on active logic* he argues for a combination of active logics and situation calculus. Time is of prime importance for other contributions as well. Mark Jago, in *Rule-based and resource-bounded: A new look at epistemic logic*, describes a modal logic of how the beliefs of agents who reason by applying rules to their existing beliefs, change over time. Michal Walicki, Marc Bezem and Wojtek Szajnkenig discuss logical aspects of state change over dense linear time, in *A strongly complete logic of dense time intervals*. Paul Egre, in *Logical omniscience and counterpart semantics*, uses counterpart semantics to model belief, while in *Impossible states at work*, Mikael Cozic extends the well known idea of using impossible worlds to model non-ideal belief to the probabilistic case, in order to model non-ideal rational choice. Aaron Kaplan, in *Simulative inference in a computational model of belief*, describes how beliefs about beliefs can come about by modeling how one reasoner can simulate the reasoning of another reasoner. In *Commitment-based decision making for bounded agents*, Olivier Roy attempts to incorporate a concept of intention, which has been argued can be useful for resource bounded decision making, into utility theory. Resource bounded agents differ not only from idealised agents, but also from each other. Fenrong Liu, in the paper *Diversity of agents*, discusses several of the ways agents can be different.

We would like to thank the workshop program committee for their excellent work in evaluating submissions to the workshop:

- Thomas Ågotnes (University of Bergen, *co-chair*)
- Natasha Alechina (University of Nottingham, *co-chair*)
- Johan van Benthem (University of Amsterdam)
- Michael Fisher (University of Liverpool)
- Chiara Ghidini (ITC-irst)
- Paul Harrenstein (Ludwig-Maximilians-Universitat Munchen)
- Wiebe van der Hoek (University of Liverpool)
- Jacek Malec (Lund University)
- Donald Perlis (University of Maryland)
- Luciano Serafini(ITC-irst)

Thomas Ågotnes and Natasha Alechina
(Workshop organisers)

Table of Contents

Belief, Knowledge, Sentences and Propositions

Rohit Parikh

Departments of Computer Science, Mathematics and Philosophy,
Brooklyn College of CUNY and CUNY Graduate Center

Abstract: Developments in reasoning about knowledge have picked up pace since the fundamental work of Jaakko Hintikka [4] and David Lewis [5]. A great deal of purely technical work has since come out of IBM [3], CUNY [7-13], Indiana, Amsterdam, and other places.

However, some philosophical issues like the justification of knowledge go back to Plato and have recently received impetus since the work of Gettier, resulting in much activity in empistemology.

Issues about belief go back at least to Frege, and the nature of belief is another active area.

Finally, starting with Aumann's work [1], there has been much activity in Game theory which is related to the issues of knowledge and common knowledge.

We will try to give a bird's eye view of some of these developments, and also say something about *what* it is that we know or believe, whether it is sentences or propositions; and if the latter, what they are. This will give us some insight into the thorny problem of logical omniscience.

Ultimately our (somewhat Wittgensteinian) view is that these notions like belief, knowledge, and common knowledge must be firmly grounded in human activity, both at the personal level (where notions like (subjective) probability, or the maximization of utility arise) and at the group level where issues of common knowledge, co-ordination, levels of knowledge, and judgment aggregation [6] arise.

References

[1] Aumann, R., Agreeing to disagree, *Annals of Statistics* **4** (1976) 1236-39.

[2] Baltag, A., and L. Moss, Logics for epistemic programs, *Synthese*, **139** (2004) 165-224

[3] Fagin, R., J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.

[4] Hintikka, J. *Knowledge and Belief*, Ithaca: Cornell University Press, 1962.

[5] Lewis, D., *Convention, a Philosophical Study*, Harvard U. Press (1969)

[6] List, C., and P. Pettit, On the many as one, *Philosophy and Public Affairs*, **33(4)** (2005) 377-390.

[7] Moss, L. and Parikh, R. Topological reasoning and the logic of knowledge, in Y. Moses (ed) *Proceedings of TARK* 1992, Morgan Kaufmann, 95-105.

[8] Pacuit, E., and R. Parikh, Social interaction, knowledge and social software, to appear in *Interactive Computation: The New Paradigm*, edited by Dina Goldin, Scott Smolka and Peter Wegner.

[9] Pacuit, E., R. Parikh and Eva Cogan, The logic of knowledge based obligation, To appear in *Knowledge, Rationality and Action*, 2006.

[10] Parikh, R., WHAT do we know and what do WE know?, in the proceedings of *Theoretical Aspects of Rationality and Knowledge*, June 2005, University of Singapore press.

[11] Parikh, R. and P. Krasucki, Communication, Consensus and Knowledge, *J. Economic Theory* **52** (1990) pp. 178-189.

[12] Parikh, R., L. Moss and C. Steinsvold, Topology and epistemic logic, to appear in a volume on *Reasoning about Space* edited by van Benthem et al.

[13] Parikh, R. and Ramanujam, R., A knowledge based semantics of messages, in *J. Logic, Language, and Information*, **12**, pp. 453 - 467, 2003.

[14] Zanaboni, Anna Maria, Notes of Rohit Parikh's lectures on Reasoning about Knowledge, (the Lectures were given in Acireale at an International School for Computer Scientists) published in Italy, summer 1993. (Cassa di Risparmio di Padova e Rovigo)

Improving Active Consequence

Mikael Asker, Jacek Malec
Department of Computer Science
Lund University

Abstract

Active consequence is a semantically defined consequence relation for active logics. It has recently been shown that active consequence is not paraconsistent, contrary to the expectations. In this paper we study a subset of the active consequence, *local consequence*, which turns out to be paraconsistent. The relation between local consequence and classical logic is investigated by first observing that the active modus ponens rule is locally unsound. But the part of the consequence relation generated by that rule which is outside local consequence is exactly that part where the premises are inconsistent. We propose the name *semantic residual* for this part, and interpret it as a reflection of artefacts in the proof theory. The residual must be included in the semantic consequence relation to achieve soundness.

1 Introduction

Active logics [EDKM$^+$99] is a family of logics intended to model awareness of resource limitations of the reasoning process. Active logics can be used for reasoning *in* time, handling contradictions and for introspection. Active logics are usually defined syntactically, as axiomatic theories, but until recently have had no formally defined semantics.

A major property of a logic capturing common intuitions about resource-awareness is *paraconsistency*. Speaking informally, this property guarantees that presence of a contradictory pair of sentences in a theory does not necessarily lead to arbitrary consequences, i.e., that the logic is not *explosive*.

The first and so far the only serious attempt to build a semantics for active logics is *active consequence* from [AGGP05b] and [AGGP05a], introduced below in Section 2. However, it has been shown in [Hov05] that active consequence is indeed explosive.

In our opinion the active consequence relation can be modified in order to escape explosiveness. In this paper we describe our attempt to constrain the active consequence so that an active logic based on it becomes paraconsistent. The rest of the paper is divided as follows: Section 2 introduces briefly active consequence. Although the paper is meant to be self-consistent, previous reading of [AGGP05b] and possibly [Hov05] may make it more legible. Section 3 introduces local consequence, a subset of active consequence with some reasonable properties, and shows that local consequence is paraconsistent. Section 4 discusses its relation to classical logic and introduces the concept of semantic residual. Then a short review of related work is presented in Section 5. Finally, Section 6 contains some conclusions and discusses how we can continue towards a complete semantics for active logics.

2 Active consequence

Active consequence was defined in [AGGP05b] as an attempt to create a semantics for a simple active logic. To make the task simple, the language is very limited. The language $\mathcal{L}$ is a sorted first-order language defined in two parts: a propositional language $\mathcal{L}_w$ to express facts about the world, and a first order language $\mathcal{L}_a$ used to express facts about the agent and its beliefs.

We will use Sn_L to denote the set of all sentences of a language L. For the languages mentioned above, this defines the sets $Sn_{\mathcal{L}}$, $Sn_{\mathcal{L}_w}$ and $Sn_{\mathcal{L}_a}$.

$\mathcal{L}_w$ is built around a set $S = \{S_i \mid i \in \mathbf{N}\}$ of sentence symbols, the propositional connectives $\neg$ and $\rightarrow$ and parentheses. The semantics of $\mathcal{L}_w$ is defined using traditional truth assignments. An $\mathcal{L}_w$-interpretation h is a function $h : Sn_{\mathcal{L}_w} \rightarrow \{\top, \bot\}$.

$\mathcal{L}_a$ is built around a set of constant symbols that can refer to points in time and to sentences in $\mathcal{L}_w$ and in $\mathcal{L}_a$ itself, the predicates now, $contra$ and bel and the propositional connective $\neg$. The intended meanings of $now(c_i)$, $contra(d_\varphi, d_{\neg\varphi}, c_i)$ and $bel(e_\varphi, c_i)$ are to indicate the current time i, the existence of a direct contradiction $\{\varphi, \neg\varphi\}$ at some time i and that the agent has a belief φ at some time i, respectively.

$\mathcal{L}_a$ is very primitive, for example it has no conjunction. As is pointed out in [Hov05], this heavily limits its expressiveness. But the focus in [AGGP05b] is on contradiction handling in the $\mathcal{L}_w$ parts of the language.

The semantics of $\mathcal{L}_a$ is based on $\mathcal{L}_a$-structures H_t^a, which describe the reasoning history of agent a at time t. An $\mathcal{L}_a$-structure H_t^a can be built from the time sequence $\langle KB_k^a \rangle_{k=0}^t$ of the knowledge bases of agent a up to time t.

The language $\mathcal{L}$ is defined so that $Sn_{\mathcal{L}} = Sn_{\mathcal{L}_w} \cup Sn_{\mathcal{L}_a}$. Its semantics is based on active structures $M_t^a = \langle h_t, H_t^a \rangle$, consisting of one $\mathcal{L}_w$-interpretation h_t and one $\mathcal{L}_a$-structure H_t^a. More details about the languages $\mathcal{L}$, $\mathcal{L}_w$ and $\mathcal{L}_a$ can be found in [Hov05].

Two kinds of consistency are defined for knowledge bases containing $\mathcal{L}$-formulas:

Definition 1 (Temporal consistency)
A set of sentences $\Sigma \subseteq Sn_{\mathcal{L}}$ is said to be

- *temporally weakly consistent at time t*, *t-weakly consistent* for short, if $\exists M_t^a : M_t^a \models (\Sigma \cap Sn_{\mathcal{L}_a})$,

- *temporally strongly consistent at time t*, *t-strongly consistent* for short, if $\exists M_t^a : M_t^a \models \Sigma$.

The sets of all finite t-weakly and t-strongly consistent subsets of $Sn_{\mathcal{L}}$ are denoted Σ_t^ω and $S_{cons\,t}$, respectively.

As in [AGGP05b], most of our databases are assumed to be t-weakly consistent.

Definition 2 (Perception function)
A *perception function* at time t is a map $per_t : \Sigma_t^\omega \rightarrow \mathscr{P}(Sn_{\mathcal{L}'})$ that is induced by an infinite sequence of non-negative integers $\langle i_1, i_2, ... \rangle$:

1. Let $\Sigma \in \Sigma_t^\omega$ and let $\Gamma = \Sigma \cap Sn_{\mathcal{L}_w}$. Order the sentences of Γ lexicographically in a string, and let $\langle S_{j_1}, S_{j_2}, ..., S_{j_n} \rangle$ be the finite sequence of all sentence-symbol tokens occurring in this string. Note that a symbol may be repeated in this sequence a number of times. Let Γ' be the set of $\mathcal{L}'_w$-sentences obtained by replacing the $\mathcal{L}_w$-symbols S_{j_k} in the $\mathcal{L}_w$-sentences in Γ with $\mathcal{L}'_w$-symbols $S_{j_k}^{i_k}$ for $1 \leq k \leq n$.

2. Let $p : \Gamma \rightarrow \Gamma'$ be the bijection mapping every sentence of Γ to its corresponding Γ'-sentence.

3. Let the set of perceived direct contradictions, denoted DC, be the set

$$\{\varphi \in \Gamma \mid p(\varphi) \in \Gamma' \text{ and } \neg p(\varphi) \in \Gamma'\}.$$

4. Finally, let $per_t(\Sigma)$ be the set

$$\begin{aligned} &[(\Sigma \setminus \Gamma) \cup \Gamma'] \\ &\setminus [\{p(\varphi) \mid \varphi \in DC\} \cup \{\neg p(\varphi) \mid \varphi \in DC\}] \\ &\cup \{contra(d_\varphi, d_{\neg\varphi}, c_t) \mid \varphi \in DC\}. \end{aligned}$$

Denote with PER_t the set of all perception functions at time t.

With the languages described, we are now ready to describe active consequence itself. The basic idea is that the agent perceives only a part of its database, and thus only some of the direct contradictions in the database are visible at a given instant.

Active consequence models this by watching the database through a *perception function*, which maps the $\mathcal{L}$-sentences in the database to another language $\mathcal{L}'$. By adding extra indexes to the sentence symbols S_j in the $\mathcal{L}_w$ formulas in the database, they are mapped to sentence symbols S_j^i in another language $\mathcal{L}'_w$.

For a direct contradiction to be visible, the sentence symbols in both of the formulas involved in the contradiction must be mapped to the same extra indexes. If they are not, then the $\mathcal{L}'_w$-symbols are different and there is no contradiction, so we can hide a direct contradiction by mapping the symbols in the $\mathcal{L}$-formulas involved to different extra indexes.

Perception functions are formally defined in Definition 2. A perception function maps a set of $\mathcal{L}$-formulas - a knowledge base - to a set of $\mathcal{L}'$-formulas - the perception of that knowledge base.

Example 1 (Perception functions) *Let Σ be the set $\{now(c_5), S_1, \neg S_2, \neg S_1\}$, and let per_5 be the perception function (at time 5) determined by the infinite superscript sequence $\langle 1, 2, 1, ...\rangle$ (only the first three elements are shown). We now wish to apply per_5 to Σ. With the terminology of Definition 2, we get $\Gamma = \{S_1, \neg S_1, \neg S_2\}$ (presented in alphabetically ordered form). Hence the finite sequence of sentence-symbol tokens is $\langle S_1, S_1, S_2\rangle$, and thus we get $\Gamma' = \{S_1^1, \neg S_1^2, \neg S_2^1\}$. Since there are no direct contradictions in Γ', we have $per_5(\Sigma) = \{now(c_5), S_1^1, \neg S_1^2, \neg S_2^1\}$.*

Should instead both occurrences of S_1 have been mapped to the same symbol, say S_1^1, we would have had a direct contradiction and the resulting image would then have been the set

$\{now(c_5), \neg S_2^1, contra(d_{S_1}, d_{\neg S_1}, c_5)\}$.

Next, we define the model set of a database of world formulas from $\mathcal{L}_w$ or $\mathcal{L}'_w$:

Definition 3 (Model set of a database of world formulas)

$$\forall\, \Sigma \in Sn_{\mathcal{L}_w} : M(\Sigma) = \{\text{All } \mathcal{L}_w\text{-interpretations } h \text{ such that } h \vDash \Sigma\}$$

$$\forall\, \Sigma \in Sn_{\mathcal{L}'_w} : M(\Sigma) = \{\text{All } \mathcal{L}'_w\text{-interpretations } h \text{ such that } h \vDash \Sigma\}$$

This will be useful later, when reasoning about the models.

We can now define active consequence itself. The basic idea is that one knowledge base follows from another, if there are ways to perceive the knowledge bases (perception functions) for which there is classical consequence between the perceptions. The following is Johan Hovold's improved (compared to the one in [AGGP05b]) definition of active consequence from [Hov05], with changed notation.

The agent index a is important, because there might be many different agents a with different histories H_{t+1}^a, but with the same databases Σ at time t and Θ at time $t+1$. [Hov05] contains a discussion about the concept of Σ-determinism, which motivates the universal quantifier for a.

The "time stamps" used by the now, $contra$ and bel predicates in this active logic are properties of the databases and are incremented by the steps taken by the consequence relation, as described in the detailed definition of $\mathcal{L}_a$-structures in [AGGP05b].

Unfortunately, in [Hov05] it was proved that active consequence is explosive. This makes it less useful as a semantic consequence relation in itself.

Definition 4 (Active consequence)
Let $\Sigma \in \Sigma_t^\omega$ be t-weakly consistent, and let $\Theta \subseteq Sn_{\mathcal{L}}$ be arbitrary. We say that Θ is a *1-step active consequence* of Σ at time t, written $\Sigma \models_{a\,1} \Theta$, if and only if

$$\begin{aligned}&\exists\, per_t \in PER_t : \exists\, per_{t+1} \in PER_{t+1} : \forall a : [KB_t^a = \Sigma \rightarrow \\ &\quad [H_{t+1}^a \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}_a}) \wedge \\ &\quad M(per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}) \neq \emptyset \wedge \\ &\quad (per_t(\Sigma) \cap Sn_{\mathcal{L}'_w}) \models (per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w})]]\end{aligned}$$

n-step active consequence is defined recursively:

$$\Sigma \models_{a\,n} \Theta \Leftrightarrow \exists\, \Gamma \subseteq Sn_{\mathcal{L}} : \Sigma \models_{a\,n-1} \Gamma \wedge \Gamma \models_{a\,1} \Theta$$

Finally, *active consequence* in general is defined as:

$$\Sigma \models_a \Theta \Leftrightarrow \exists\, n \in \mathbf{N} : \Sigma \models_{a\,n} \Theta$$

3 Local consequence

The informal reasoning behind active consequence in [AGGP05b] is that the agent only perceives a part of its knowledge base and thus only a part of the contradictions in it. We find that reasoning appealing and our working hypothesis has been that the authors of [AGGP05b] did good informal background reasoning, and just made some formal mistakes.

An analysis of the explosivity proof in [Hov05] shows that one important problem is that we can map the same variable to two values at the same time. This is what causes the explosion, and in this sense, the perceptions functions in active consequence are too general.

One idea about how to cure the problems is to restrict the choice of perception functions. In some active logics systems, such as the memory model in [DMP86] and the $\mathbb{L}_{mm}$ system in [Ask03], the subdivision of the knowledge base is done at the formula level, not at the term level. So one reasonable way to restrict the choice of perception functions is to take subsets of the formulas at the left-hand side (lhs) and right-hand side (rhs). This also eliminates the "two values at the same time" problem. If we require the subsets to be consistent, then classical logical consequence between the perceptions is meaningful in the sense that models do exist.

If we then have a consistent subset on the lhs, what should follow from it on the rhs? Classical logic consequence requires that the models of the lhs are a subset of the models of the rhs. This means that new models can be added, but old models can't be deleted. The new formulas added to the rhs might possibly restrict the model set, which is what we want to avoid. We may avoid it if the new formulas follow classically from the subset. The old formulas on the rhs can't add any new restrictions on the models, because they are also present on the lhs.

So what we want to do is to take a consistent subset of the lhs and take the subset of the rhs which contains the new formulas. Technically, this can be realized inside the framework of active consequence, by mapping the symbols in a formula to index 1 iff the formula is a member of the subset. Otherwise its terms are mapped to unique indexes different from 1 to avoid restricting the model set. One more requirement which turns out to be necessary is that the old formulas on the right side are mapped to the same indexes as on the left side. These restrictions on the perception functions are expressed formally in the following definition.

Definition 5 (Local perception function pair)
Let $\Sigma \in \Sigma_t^\omega$ and $\Theta \subseteq Sn_{\mathcal{L}}$ and denote with Γ_1 the set $\Sigma \cap Sn_{\mathcal{L}_w}$ and with Γ_2 the set $\Theta \cap Sn_{\mathcal{L}_w}$. Order the sentences in Γ_1 and Γ_2 lexicographically as $\varphi_{11}, \varphi_{12}, ..., \varphi_{1m_1}$ and $\varphi_{21}, \varphi_{22}, ..., \varphi_{2m_2}$, respectively. Let $\langle S_{j_{11}}, S_{j_{12}}, ..., S_{j_{1n_1}} \rangle$ and $\langle S_{j_{21}}, S_{j_{22}}, ..., S_{j_{2n_2}} \rangle$ be the finite sequences of all sentence-symbol tokens occurring in the strings $\varphi_{11}\varphi_{12}...\varphi_{1m_1}$ and $\varphi_{21}\varphi_{22}...\varphi_{2m_2}$, the concatenated formulas. For each k_1 in $\{1, ..., n_1\}$ or k_2 in $\{1, ..., n_2\}$, there is a corresponding formula φ_{1l}, $1 \leq l \leq m_1$ or φ_{2l}, $1 \leq l \leq m_2$, to which the sentence symbol $S_{j_{1k_1}}$ or $S_{j_{2k_2}}$ belongs.

For each $\Delta \subseteq \Gamma_1$, we can define a sequence of positive integers $\langle i_{11}, i_{12}, ... \rangle$ by

$$i_{1k} = \begin{cases} 1 & \text{if } \varphi_{1l} \in \Delta, \text{ where } \varphi_{1l} \text{ is} \\ & \quad \text{the formula in } \Gamma_1 \text{ containing } S_{j_{1k}} \\ & \text{or if } k > n_1 \\ k+1 & \text{otherwise} \end{cases}$$

We also define the sequence $\langle i_{21}, i_{22}, ... \rangle$ of positive integers as

$$i_{2k} = \begin{cases} 1 & \text{if } \varphi_{2l} \in \Gamma_2 \setminus \Gamma_1, \text{ where } \varphi_{2l} \text{ is} \\ & \quad \text{the formula in } \Gamma_2 \text{ containing } S_{j_{2k}} \\ & \text{or if } k > n_2 \\ \text{otherwise } \varphi_{2l} \in \Gamma_2 \cap \Gamma_1 \subseteq \Gamma_1 & \text{- use the same value} \\ \quad \text{as that used for } \varphi_{2l} \text{ in the } i_{1k} & \text{sequence} \end{cases}$$

The sequences $\langle i_{11}, i_{12}, ... \rangle$ and $\langle i_{21}, i_{22}, ... \rangle$ above define together a pair of perception functions (per_t, per_{t+1}).

The set $PER_{l\,t}(\Sigma, \Theta)$ is the set of perception function pairs defined as above for all *consistent* subsets $\Delta \subseteq \Sigma \cap Sn_{\mathcal{L}_w}$. This defines the function

$$PER_{l\,t} : \Sigma_t^\omega \times \mathscr{P}(Sn_{\mathcal{L}}) \rightarrow \mathscr{P}(PER_t) \times \mathscr{P}(PER_{t+1})$$

which generates the set of possible *local perception function pairs* from the sets of premises and results.

We can now use the $PER_{l\,t}$ restriction on the perception functions to define a new consequence relation:

Definition 6 (Local consequence)
The consequence relation $\vDash_{l\,1}$ is defined by restricting per_t and per_{t+1} with $PER_{l\,t}$ in the definition of $\vDash_{a\,1}$. The consequence relation $\vDash_{l\,n}$ is defined recursively from $\vDash_{l\,1}$, in the same way as $\vDash_{a\,n}$ is defined from $\vDash_{a\,1}$, and the relation $\vDash_l$ is defined from $\vDash_{l\,n}$ in the same way as $\vDash_a$ is defined from $\vDash_{a\,n}$.

The term local is loosely derived from the idea that locally, between the visible consistent parts of the databases, there is classical consequence.

The following theorem shows that Definition 6 is correct, in the sense that the new consequence relation has the desired properties:

Theorem 1 (Local consequence is local)

$$\forall\, (\Sigma, \Theta) \in \Sigma_t^\omega \times \mathscr{P}(Sn_{\mathcal{L}}) :$$
$$\Sigma \models_{l\,1} \Theta \Rightarrow \exists\, \Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} :\, M(\Sigma') \neq \emptyset \,\wedge \Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w}$$

$$\forall\, (\Sigma, \Theta) \in \mathscr{P}(Sn_{\mathcal{L}_w}) \times \mathscr{P}(Sn_{\mathcal{L}_w}) :$$
$$(\exists\, \Sigma' \subseteq \Sigma :\, M(\Sigma') \neq \emptyset \,\wedge \Sigma' \models \Theta \setminus \Sigma) \Rightarrow \Sigma \models_{l\,1} \Theta$$

Proof: See Appendix A.

Next, we show that local consequence is a subset of active consequence:

Theorem 2 (Local consequence is active) $\models_l \subset \models_a$.

Proof:

$$\forall\, (\Sigma, \Theta) \in \Sigma_t^\omega \times \mathscr{P}(Sn_{\mathcal{L}}) :\, PER_{l\,t}(\Sigma, \Theta) \subset PER_t \times PER_{t+1} \tag{1}$$

So the set of perception functions used when defining $\models_{l\,1}$ is a subset of the set of perception functions used when defining $\models_{a\,1}$.

□

This property is used in Section 4 when investigating the relation between local consequence and classical logic.

In [Hov05], active consequence is said to be explosive because the explosive rule

$$\frac{t : \varphi, \neg\varphi}{t+1 : \psi}, \quad \text{where } \varphi, \psi \in Sn_{\mathcal{L}_w} \tag{2}$$

is active sound. By analogy, local consequence would be explosive if

$$\forall\, \Sigma \in \Sigma_t^\omega : \forall\, (\varphi, \psi) \in Sn_{\mathcal{L}_w} \times Sn_{\mathcal{L}_w} : \{\varphi, \neg\varphi\} \subseteq \Sigma \Rightarrow \Sigma \models_l \Sigma \cup \{\psi\} \tag{3}$$

and otherwise paraconsistent. We will now show that it in fact *is* paraconsistent:

Theorem 3 (Local consequence is paraconsistent)

$$\neg\, \forall\, \Sigma \in \Sigma_t^\omega : \forall\, (\varphi, \psi) \in Sn^2_{\mathcal{L}_w} : \{\varphi, \neg\varphi\} \subseteq \Sigma \Rightarrow \Sigma \models_l \Sigma \cup \{\psi\}$$

Proof: See Appendix B.

4 Relation to classical logic

In Section 3, we claimed that $\models_l$ is paraconsistent. We will now show that the world part of local consequence is classical, a property of $\models_l$ which is used when proving its paraconsistency in Appendix B:

Theorem 4 (The world part of local consequence is classical)

$$\forall\, (\Sigma, \Theta) \in \Sigma_t^\omega \times \mathscr{P}(Sn_{\mathcal{L}}) :\, \Sigma \models_{l\,1} \Theta \Rightarrow \Sigma \cap Sn_{\mathcal{L}_w} \models \Theta \cap Sn_{\mathcal{L}_w}$$

Proof:

If $\Sigma \models_{l\,1} \Theta$, then by Theorem 1 we have

$$\exists\, \Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} :\ M(\Sigma') \neq \emptyset\ \wedge \Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \tag{4}$$

We have

$$\Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} \Rightarrow \Sigma \cap Sn_{\mathcal{L}_w} \models \Sigma' \tag{5}$$

and we get

$$\begin{array}{l} \Sigma \cap Sn_{\mathcal{L}_w} \models \Sigma' \wedge \Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \Rightarrow \\ \Sigma \cap Sn_{\mathcal{L}_w} \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \Rightarrow \\ \Sigma \cap Sn_{\mathcal{L}_w} \models \Theta \cap Sn_{\mathcal{L}_w} \end{array} \tag{6}$$

□

Next, we show that for strongly consistent databases, $\models_l$ has the same nice relation to classical logic as that proved in [Hov05] for $\models_a$:

Theorem 5 *For consistent world formulas, local consequence is equivalent to classical consequence:*

$$\forall\ (\Sigma, \Theta) \in [\mathscr{P}(Sn_{\mathcal{L}_w})]^2 : M(\Sigma) \neq \emptyset \Rightarrow (\Sigma \models_{l\,1} \Theta \Leftrightarrow \Sigma \models \Theta)$$

Proof:

$\models_{l\,1} \Rightarrow \models$) : Use Theorem 4 above.

$\models_{l\,1} \Leftarrow \models$) :

We have

$$\Sigma \models \Theta \Rightarrow \exists\, \Sigma' \subseteq \Sigma : \Sigma' \models \Theta \tag{7}$$

and

$$\Sigma' \models \Theta \Rightarrow \Sigma' \models \Theta \setminus \Sigma \tag{8}$$

So if $M(\Sigma) \neq \emptyset$, we get

$$\exists\, \Sigma' \subseteq \Sigma : M(\Sigma') \neq \emptyset \wedge \Sigma' \models \Theta \setminus \Sigma \tag{9}$$

which by Theorem 1 implies that $\Sigma \models_{l\,1} \Theta$.

□

Let $\vdash_{amp}$ be the consequence relation which is generated by the *active modus ponens* inference rule, the active logics variant of the traditional modus ponens rule:

$$\frac{t : \varphi, \varphi \rightarrow \psi}{t+1 : \psi}, \quad \text{where } \varphi, \psi \in Sn_{\mathcal{L}_w} \tag{10}$$

Active modus ponens is certainly a rule which we want to be local sound. But unfortunately, because

$$\{\neg(S_1 \rightarrow S_1), (\neg(S_1 \rightarrow S_1)) \rightarrow S_2\} \nvDash_l \{S_2\} \tag{11}$$

it isn't, $\vdash_{amp} \setminus \models_l \neq \emptyset$!

In the proof of active soundness for the active modus ponens rule in [AGGP05b], all variables in the premises are mapped to different indexes, to guarantee that the premises are consistent and have models. That can't be done here, so in this sense local consequence is a too restrictive subset of active consequence.

We want to handle active modus ponens, but still be paraconsistent. So we have to find a consequence relation $\models_x$ "between" $\models_l$ and $\models_a$, with $\models_l \subset \models_x$ to include active modus ponens, but with $\models_x \subset \models_a$ to remain paraconsistent.

The part of $\vdash_{amp}$ that is inside $\models_l$ is the part where the premises have models and the reasoning is "truly meaningful". The part of $\vdash_{amp}$ that is outside $\models_l$, the *semantic residual*, is exactly the part where the premises are inconsistent. We interpret the residual as a reflection of artifacts in the proof theory. The residual models the proof theory when it is doing "meaningless" reasoning with inconsistent premises.

We believe that there is a reason for why the proof theory does this kind of "meaningless" reasoning, and that the reason is that the inference rule has no way to find out that the reasoning is "meaningless". Inference rules are syntactic operations which can be implemented efficiently with simple pattern matching, and are "too stupid" to discover inconsistency. Computationally, discovering inconsistency is expensive. The residual part is present also in classical logic, it is just that nobody has thought of it before. The behavior of classical logic with inconsistent databases is so ugly that nobody has bothered with the details.

The residual part must be included in a semantic consequence relation to achieve soundness, and because active modus ponens is *active* sound, we can include the residual in the semantic consequence relation and still have that relation being a subset of active consequence. Now the utility of being a subset of active consequence as proved in Theorem 2 becomes apparent. That property enables us to go outside $\models_l$ to other parts of $\models_a$ to cover the residual parts of the rules.

5 Related work

The work described in this paper directly builds on the earlier work on Active Logics: logics aware of passage of time and the computational cost of inference. An early attempt in this direction has been reported in [Lev84], further improved and extended in several ways by [FH88]. However, both approaches suffer from the (possibly partial) omniscience problem: the systems cannot model agents with finite memory resources. The next steps, attempting to overcome this issue, were the Memory Model [DMP86] and the step-logic [EDP88] that evolved into a family of *active logics* [EDKM$^+$99]. Active logics have been used so far to describe a variety of domains, however, none of the proposed systems has been shown to overcome the limitation of the exponential blow-up of the number of formulae produced in the inference process. Another property of active logics is that until very recently they missed well-defined semantics.

The first effort to provide a natural semantics for an active logic system has been reported by [ED88], but the semantics developed there was insufficient to characterize the logic properly. Another attempt to address this complex issue has not been done until very recently [AGGP05b], where active consequence relation has been introduced. Unfortunately, even this characterisation was not completely correct, as it has been shown in [Hov05]. Yet another possibility, to embed active logics in labelled deductive systems, has been reported in [AM05], but without any conclusive results. We are not aware of any other semantics for an active logic or a similar system.

On a slightly larger perspective, active logics are one of the attempts to address the question of non-omniscience of real agents. There exists a number of approaches that try to deal with this problem. Speaking generally, any such solution must address the need to model bounded resources of agents and, independently, its incomplete reasoning mechanisms.

The approach of Fagin, Halpern, Moses and Vardi [FHMV03], *Interpreted multi-agent systems* captures in a nice way the evolution of an agent's knowledge. The system consists of the usual knowledge/belief modalities mixed with the classical temporal operators. The resulting system is interpreted on so called *runs*. The system is too powerful for our needs, although it may be modified in the direction of non-omniscient agents. The same remarks apply to the recent proposal of van der Hoek and Wooldridge, *Alternating-time Temporal Epistemic Logic (ATEL)*, [vdHW03], in which the usual knowledge/belief modalities (K_x) and the classical temporal operators are extended with a dynamic-logic-like concept of *cooperative actions*. The interpretations are based on *concurrent game structures*. Although the authors mention the possibility of describing non-omniscient agents, the main system is developed for at least partially omniscient entities. A similar system, based on active logic, has been proposed by Grant, Kraus and Perlis [GKP00].

The last two systems we would like to mention in this context are rather similar. The first is *Timed Reasoning Logics* by Alechina et al. [ALW04], the second *Logic of Finite Syntactic Epistemic States* proposed recently by Ågotnes [Ågo04]. His system is based on ATEL, but does not assume any structure in the underlying language — the epistemic states of an agent are purely syntactical structures. Knowledge

evolution mechanisms are modeled using *rules*; they need not to be necessarily sound nor complete. Although appealing from the formal point of view, this system does not provide any hints about dealing with the computational complexity of the problem.

There is a growing insight that logic, should it be considered as a useful tool for building autonomous intelligent agents, has to be used in a substantially different way than before. Active logics are one example of this insight, while other important contributions might be found, e.g., in [GW01] or [WL01].

6 Conclusions

Active consequence in its original form is not very useful as a semantics for active logics itself, because it is explosive. But this text shows that it is indirectly useful, in that can be used as a framework for defining other consequence relations which are subsets of active consequence, by restricting the choice of perception functions.

One such subset is local consequence, which was defined and shown to be paraconsistent in Section 3. As explained there, we probably want the world part of the reasoning to follow local consequence.

The following figure shows the relationships between some interesting consequence relations:

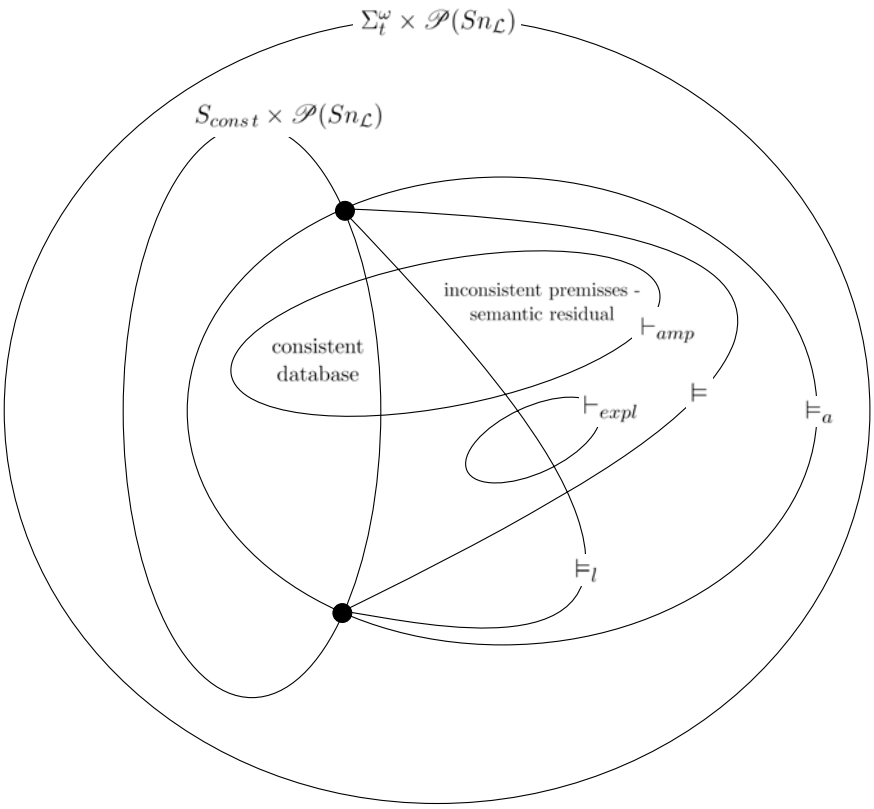


Figure 1: Some relationships between the different consequence relations. $\Sigma_t^\omega \times \mathscr{P}(Sn_\mathcal{L})$ is the set of all possible relations between pairs of finite t-weakly consistent databases. S_{const} is the subset of all databases in Σ_t^ω which are t-strongly consistent, so $S_{const} \times \mathscr{P}(Sn_\mathcal{L})$ is the subset of $\Sigma_t^\omega \times \mathscr{P}(Sn_\mathcal{L})$ where the original database is strongly consistent. $\vdash_{amp}$ is the consequence relation generated by the active modus ponens rule, $\vdash_{expl}$ is the consequence relation generated by the explosive rule, $\models$ is classical consequence, $\models_l$ is local consequence and $\models_a$ is active consequence.

As shown in Section 4, the active modus ponens inference rule is locally unsound, and the part of $\vdash_{amp}$ which is outside $\models_l$ is exactly the part where the premises are inconsistent. And as explained in Section 4, we interpret this as the inference rule being too "stupid" to realize that there are no models. By adding the

semantic residual part to local consequence, we get a new consequence relation which covers active modus ponens but still is paraconsistent. Hopefully, this procedure can be generalized to other world part inference rules.

So what we want as our final semantic consequence relation is probably a part of local consequence, plus some semantic residuals to compensate for the stupidity of our inference rules.

7 Acknowledgements

Many thanks to Michael L. Anderson, Walid Gomaa, John Grant and Don Perlis for inventing active consequence and to Johan Hovold for analyzing and improving it. Their contributions provide the foundation of this work. We also thank Sven Asker for proof-reading the manuscript. Finally Mikael wants to thank the dog Basker, which recently passed away.

References

[AGGP05a] Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis. Active logic semantics for a single agent in a static world. Manuscript from October 14, 2005, 2005.

[AGGP05b] Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis. On the reasoning of real-world agents: Toward a semantics for active logic. In *Proceedings of the 7th Annual Symposium on the Logical Formalization of Commonsense Reasoning*, Corfu, Greece, 2005.

[Ågo04] T. Ågotnes. *A Logic of Finite Syntactic Epistemic States*. PhD thesis, Department of Informatics, University of Bergen, Norway, 2004.

[ALW04] N. Alechina, B. Logan, and M. Whitsey. A complete and decidable logic for resource-bounded agents. In *Proc. AAMAS'04*, 2004.

[AM05] Mikael Asker and Jacek Malec. Reasoning with limited resources: active logics expressed as labelled deductive systems. *Bulletin of the Polish Academy of Sciences*, 53(1):69–78, 2005.

[Ask03] Mikael Asker. Logical reasoning with temporal constraints. Master's thesis, Department of Computer Science, Lund University, Lund, Sweden, 2003. Available at http://ai.cs.lth.se/xj/MikaelAsker/exjobb0820.ps.

[DMP86] Jennifer Drapkin, Michael Miller, and Donald Perlis. A memory model for real-time commonsense reasoning. Technical Report TR-86-21, Systems Research Center, University of Maryland, College Park, Maryland, 1986.

[ED88] Jennifer Elgot-Drapkin. Step-logic: Reasoning situated in time. PhD thesis CS-TR-2156, Department of Computer Science, University of Maryland, College Park, Maryland, 1988.

[EDKM+99] Jennifer Elgot-Drapkin, Sarit Kraus, Michael Miller, Madhura Nirkhe, and Donald Perlis. Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-4072, Department of Computer Science, University of Maryland, College Park, Maryland, 1999.

[EDP88] Jennifer Elgot-Drapkin and Donald Perlis. Reasoning situated in time I: Basic concepts. Technical Report CS-TR-2016, Department of Computer Science, University of Maryland, College Park, Maryland, April 1988.

[FH88] R. Fagin and J.Ỹ. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[FHMV03] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 2003.

[GKP00] J. Grant, S. Kraus, and D. Perlis. A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems*, 3(4):455–458, 2000.

[GW01] D. Gabbay and J. Woods. The new logic. *L. J. of the IGPL*, 9(2):141–174, 2001.

[Hov05] Johan Hovold. On a semantics for active logic. Master's thesis, Department of Computer Science, Lund University, Lund, Sweden, 2005. Available at http://ai.cs.lth.se/xj/JohanHovold/finalreport.pdf.

[Lev84] H. Levesque. A logic of implicit and explicit belief. In *Proc. AAAI 84*, pages 198–202, 1984.

[vdHW03] W. van der Hoek and M. Wooldridge. Cooperation, knowledge and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75:125–157, 2003.

[WL01] M. Wooldridge and A. Lomuscio. A computationally grounded logic of visibility, perception, and knowledge. *L. J. of the IGPL*, 9(2):257–272, 2001.

A Correctness proof

First we need some help definitions:

Definition 7 (Mapping functions)
Let $\Sigma \in \Sigma_t^\omega$, $\Theta \subseteq Sn_{\mathcal{L}}$, $(per_t, per_{t+1}) \in PER_{l\,t}(\Sigma, \Theta)$ and let $\Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w}$ be the consistent subset used for (per_t, per_{t+1}) in the $PER_{l\,t}$ definition.

per_t maps the formulas in $\Sigma \cap Sn_{\mathcal{L}_w}$ to $\mathcal{L}'_w$-formulas, the formulas in Σ' using index 1 and the other formulas using unique indexes $\neq 1$. For a given $\Sigma'' \subseteq \Sigma'$, we let $map_1(\Sigma'')$ denote the $\mathcal{L}'_w$-formulas made by per_t from Σ'', and for a given $\Sigma'' \subseteq (\Sigma \cap Sn_{\mathcal{L}_w}) \setminus \Sigma'$, we let $map_{unique}(\Sigma'')$ denote the same thing.

For per_{t+1}, we define map_1 and map_{same} in the same manner.

Proof of Theorem 1:

Part 1 : $\models_{l\,1} \Rightarrow \models$

We want to prove that

$$\begin{array}{l} \forall\ (\Sigma, \Theta) \in \Sigma_t^\omega \times \mathscr{P}(Sn_{\mathcal{L}}) : \\ \quad \Sigma \models_{l\,1} \Theta \Rightarrow \\ \quad\quad \exists\ \Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} :\ M(\Sigma') \neq \emptyset\ \wedge \Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \end{array} \tag{12}$$

By the definition of $\models_{l\,1}$ we have

$$\begin{array}{l} \Sigma \models_{l\,1} \Theta \Rightarrow \\ \quad \exists\ (per_t, per_{t+1}) \in PER_{l\,t}(\Sigma, \Theta) : \\ \quad\quad per_t(\Sigma) \cap Sn_{\mathcal{L}'_w} \models per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w} \end{array} \tag{13}$$

Because $(per_t, per_{t+1}) \in PER_{l\,t}(\Sigma, \Theta)$, there exists a consistent subset $\Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w}$ of Σ which together with Σ and Θ generates the number sequences which generates the perception functions according to the definition of $PER_{l\,t}$. Because of the way the sequences are constructed, we get

$$\begin{array}{l} map_1(\Sigma') \cup map_{unique}((\Sigma \setminus \Sigma') \cap Sn_{\mathcal{L}_w}) \models \\ map_1((\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w}) \cup map_{same}((\Theta \cap \Sigma) \cap Sn_{\mathcal{L}_w}) \end{array} \tag{14}$$

Because reducing a set of formulas enlarges its set of models, we can remove $map_{same}((\Theta \cap \Sigma) \cap Sn_{\mathcal{L}_w})$ from the right side of the $\models$-expression:

$$map_1(\Sigma') \cup map_{unique}((\Sigma \setminus \Sigma') \cap Sn_{\mathcal{L}_w}) \models map_1((\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w}) \tag{15}$$

Because there are no restrictions on variables with index other than one on the right side, we can safely remove all such restrictions from the left side :

$$map_1(\Sigma') \models map_1((\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w}) \tag{16}$$

Now we can remove the indexes and take the step from $Sn_{\mathcal{L}'_w}$ to $Sn_{\mathcal{L}_w}$:

$$\Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \tag{17}$$

Part 2 : $\models \Rightarrow \models_{l\,1}$

We want to prove that

$$\begin{aligned} &\forall\, (\Sigma, \Theta) \in [\mathscr{P}(Sn_{\mathcal{L}_w})]^2 : \\ &\quad \exists\, \Sigma' \subseteq \Sigma :\ M(\Sigma') \neq \emptyset\ \wedge \Sigma' \models \Theta \setminus \Sigma \Rightarrow \\ &\qquad \Sigma \models_{l\,1} \Theta \end{aligned} \tag{18}$$

If $\Sigma' \subseteq \Sigma \wedge M(\Sigma') \neq \emptyset$, then by the definition of $PER_{l\,t}$, Σ, Σ' and Θ together generate two number sequences which generate a local perception function pair $(per_t, per_{t+1}) \in PER_{l\,t}(\Sigma, \Theta)$.

If we have

$$\Sigma' \models \Theta \setminus \Sigma \tag{19}$$

then we can add indexes and go from $Sn_{\mathcal{L}_w}$ to $Sn_{\mathcal{L}'_w}$:

$$map_1(\Sigma') \models map_1(\Theta \setminus \Sigma) \tag{20}$$

Because enlarging a set of formulas reduces its set of models, we can add $map_{unique}(\Sigma \setminus \Sigma')$ to the left side of the $\models$-expression:

$$map_1(\Sigma') \cup map_{unique}(\Sigma \setminus \Sigma') \models map_1(\Theta \setminus \Sigma) \tag{21}$$

By using $\Sigma' = (\Sigma' \setminus \Theta) \cup (\Sigma' \cap \Theta)$ and $\Sigma \setminus \Sigma' = [(\Sigma \setminus \Sigma') \setminus \Theta] \cup [(\Sigma \setminus \Sigma') \cap \Theta]$, we get

$$\begin{aligned} &map_1(\Sigma' \setminus \Theta) \cup map_1(\Sigma' \cap \Theta) \cup \\ &map_{unique}((\Sigma \setminus \Sigma') \setminus \Theta) \cup map_{unique}((\Sigma \setminus \Sigma') \cap \Theta) \models \\ &\quad map_1(\Theta \setminus \Sigma) \end{aligned} \tag{22}$$

We can add formulas from the left side of a $\models$-expression to the right side :

$$\begin{aligned} &map_1(\Sigma' \setminus \Theta) \cup map_1(\Sigma' \cap \Theta) \cup \\ &map_{unique}((\Sigma \setminus \Sigma') \setminus \Theta) \cup map_{unique}((\Sigma \setminus \Sigma') \cap \Theta) \models \\ &\quad map_1(\Sigma' \cap \Theta) \cup map_{unique}((\Sigma \setminus \Sigma') \cap \Theta) \cup map_1(\Theta \setminus \Sigma) \end{aligned} \tag{23}$$

By using $\Sigma \cap \Theta = (\Sigma' \cap \Theta) \cup [(\Sigma \setminus \Sigma') \cap \Theta]$ and rewriting the left side back to its original form, we get

$$map_1(\Sigma') \cup map_{unique}(\Sigma \setminus \Sigma') \models map_1(\Theta \setminus \Sigma) \cup map_{same}(\Theta \cap \Sigma) \tag{24}$$

By the definition of $PER_{l\,t}$, that is the same as

$$per_t(\Sigma) \cap Sn_{\mathcal{L}'_w} \models per_{t+1}(\Theta) \cap Sn_{\mathcal{L}'_w} \tag{25}$$

and this means that by the definition of $\models_{l\,1}$

$$\Sigma \models_{l\,1} \Theta \tag{26}$$

□

B Paraconsistency proof

Proof of Theorem 3:

By the definition of paraconsistency, $\models_l$ is paraconsistent if

$$\exists\, (\varphi, \psi) \in {Sn_{\mathcal{L}_w}}^2 : \{\varphi, \neg\varphi\} \not\models_l \{\psi\} \tag{27}$$

So we can prove paraconsistency by choosing $\varphi = S_1$, $\Sigma = \{\varphi, \neg\varphi\} = \{S_1, \neg S_1\}$, $\psi = S_2$ and $\Theta = \{\psi\} = \{S_2\}$ and proving that

$$\Sigma \not\models_l \Theta \tag{28}$$

which by the definition of $\models_l$ is equivalent to

$$\forall n \in \mathbf{N} :\ \Sigma \not\models_{l\, n} \Theta \tag{29}$$

We have two cases :

Case 1 : $n = 1$

$\Sigma = \{S_1, \neg S_1\}$ has only three consistent subsets: $\emptyset$, $\{S_1\}$ and $\{\neg S_1\}$. Because none of these subsets contain any restriction of the variable S_2, all of these subsets have models in which S_2 is false.

But in all models of $\Theta \setminus \Sigma = \{S_2\} \setminus \{S_1, \neg S_1\} = \{S_2\}$, the variable S_2 is true. This means that

$$\forall\, \Sigma' \subseteq \Sigma : M(\Sigma') \neq \emptyset \Rightarrow M(\Sigma') \setminus M(\Theta \setminus \Sigma) \neq \emptyset \Rightarrow \Sigma' \not\models \Theta \setminus \Sigma \tag{30}$$

which means that

$$\neg\, \exists\, \Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} : M(\Sigma') \neq \emptyset \wedge \Sigma' \models (\Theta \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \tag{31}$$

which by Theorem 1 means that

$$\Sigma \not\models_{l\, 1} \Theta \tag{32}$$

So (29) is true for $n = 1$.

Case 2 : $n > 1$

Assume that (29) is false for n:

$$\Sigma \models_{l\, n} \Theta \tag{33}$$

By the definition of $\models_{l\, n}$, because $n > 1$ this is the same as

$$\exists\, \Gamma \subseteq Sn_{\mathcal{L}} :\ \Sigma \models_{l\, 1} \Gamma \wedge \Gamma \models_{l\, n-1} \Theta \tag{34}$$

By Theorem 1, we have

$$\Sigma \models_{l\, 1} \Gamma \Rightarrow \exists\, \Sigma' \subseteq \Sigma \cap Sn_{\mathcal{L}_w} : M(\Sigma') \neq \emptyset \wedge \Sigma' \models (\Gamma \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \tag{35}$$

Because $\Sigma \cap Sn_{\mathcal{L}_w} = \Sigma = \{S_1, \neg S_1\}$ contains nothing except for a direct contradiction, all of it will be removed by the direct contradiction removal mechanism in $\Sigma \models_{l\, 1} \Gamma$, and we have $\Sigma \cap \Gamma = \emptyset$ and $\Gamma \cap Sn_{\mathcal{L}_w} = (\Gamma \setminus \Sigma) \cap Sn_{\mathcal{L}_w}$.

So from

$$\Sigma' \models (\Gamma \setminus \Sigma) \cap Sn_{\mathcal{L}_w} \tag{36}$$

we get

$$\Sigma' \models \Gamma \cap Sn_{\mathcal{L}_w} \Leftrightarrow M(\Sigma') \subseteq M(\Gamma \cap Sn_{\mathcal{L}_w}) \tag{37}$$

Above we found that Σ' has a model in which S_2 is false, so that must also apply to $\Gamma \cap Sn_{\mathcal{L}_w}$.

Because of Theorem 4, we have

$$\begin{array}{l} \Sigma \models_{l\,1} \Theta \Rightarrow \\ \Sigma \cap Sn_{\mathcal{L}_w} \models \Theta \cap Sn_{\mathcal{L}_w} \Rightarrow \\ M(\Sigma \cap Sn_{\mathcal{L}_w}) \subseteq M(\Theta \cap Sn_{\mathcal{L}_w}) \end{array} \tag{38}$$

and because of the transitivity of $\subseteq$, we get

$$\Sigma \models_{l\,n-1} \Theta \Rightarrow M(\Sigma \cap Sn_{\mathcal{L}_w}) \subseteq M(\Theta \cap Sn_{\mathcal{L}_w}) \tag{39}$$

So because $\Gamma \cap Sn_{\mathcal{L}_w}$ has a model in which S_2 is false, then that must also apply to $\Theta' \cap Sn_{\mathcal{L}_w}$. This contradicts $\Theta = \{S_2\} \subseteq \Theta'$, so assumption (33) was wrong and (29) is true for all $n > 1$.

□

A modal logic of time-bounded reasoning agents

Ayelet Butman[1] and Sarit Kraus[2]

[1] Holon Institute of technology, Holon, Israel
ayeletb@hit.ac.il
[2] Bar-Ilan University, Ramat-Gan, Israel
sarit@cs.biu.ac.il

Abstract. Real agents must work within the limitations or *bounds* imposed by their environment and their own makeup. Among those resources available only in limited quantities are time, space and explicit information. Most formal attempts to model agents assume that an agent is able to reason forever in a timeless present as if the world had stopped for the agent's benefit. This work is intended to narrow the gap between formal models of agents and realistic agents. We consider the issue of limitation of reasoning. The semantics we use are based upon Montague semantics.

1 Introduction

All agents which function in the real world, whether human or automated, are subject to the passage of time during the reasoning process. There are numerous problems in AI-planning and common-sense reasoning in which the capacity to reason and act *in* time is of paramount importance. Consider, for example, the following problem in which the passage of time (while the agent reasons) is crucial (taking from [Nir94]):
Examination problem: A student taking an examination must determine a strategy for deciding on which problems to work, how much time to allocate for each, etc. Yet every second spent in such decision-making is a second less for problem-solving. Deliberation time is a significant amount of the total time available - time which must be factored into the reasoning.

Most formal and commonsense reasoning approaches to modelling agents do not have an appropriate representational framework for tackling time-situated reasoning problem such as this above. They assume that an agent reasons forever in a timeless present as if the world had stopped for the agent's benefit. Resource limitations for modeling reasoning agents have been of some concern in formal logical work. In particular, the problem of *logical omniscience* has received attention in the epistemic logic literature (e.g., [GFV05,FHV95,FH88,Lev84,LL00]). Work in temporal logic involves reasoning *about* time (e.g., [All84,McD82]). However, in temporal logic time is not treated as a crucial resource that must be carefully rationed by the agent.

The syntactic approaches has been suggested by Konolige([Kon83], and an influential recent approach to the problem of bounded reasoning agents uses the

framework of Gabbay's Labelled Deductive Systems (LDS) by Alechina, Logan and Whitsey ([Ale04].

Formal system representing reasoning an agent's ongoing process of reasoning have been use for example in [NKPM97]. One of the important aspect of omniscience logical problem is the limitation ability of artificial intelligence agents to conclude. This paper focus on this problem.

In this study[3] unlike the above mentioned approaches we pursue in this paper a purely semantical approach to resource bounded reasoning.

Montague has given a possible world semantics that partially avoids the problem of logical omniscience. We use the main idea of Montague's model, namely to define knowledge as a relation between a world and a set of sets of possible worlds. However, we provide the distinction of incorporating time to the model. Vardi [Var86] provides a co-relationship between restrictions on models in the Montague semantics and the corresponding agent properties which they characterize. We apply his results to our framework.

Fagin and Halpern [FH88] have presented a series of interesting approaches to limited reasoning which merge the syntactic and semantic approaches. They provide an extension to Levesque's approach for the multi-agent case, and introduce a notion of *awareness*. They also provide, what they refer to as a *society of minds approach* approach to local reasoning. Fagin and Halpern's awareness notion, in their logic of general awareness, acts like a filter on semantic formulations. A problematic feature of this model has been that an agent can compute all logical consequences of which it is aware, given memory limitations.

There are a number of works which consider logics of knowledge and time e.g., [KL88]. Fagin and Halpern discuss the possibility of capturing bounded and situated reasoning by letting the awareness set vary over time. We will attempt a systematic modeling of situations where the passage of time is a critical issue.

2 A time bounded approach to reasoning

The sources of our model is the approach of Active-logics, see e.g. [EDP90]. In our modal A *step* is defined as a fundamental unit of inference time. Beliefs are parameterized by the time taken for their inference, and these time parameters can themselves play a role in the specification of the inference rules and axioms. The most obvious way time parameters can enter is via the expression $Now(i)$, indicating that the time is now i. Each step of reasoning advances i by 1. At each new step i, the only information available to the agent upon which to base further reasoning is a snap-shot of his deduction process completed up to and including step $i-1$.

One of the main shortcomings of the original active-logics is the unrealistic parallelism: it was assumed that during a given step i, the agent can apply all available inference rules in parallel to the beliefs at step $i-1$. A short-term memory was introduced in the formalism of active-logic to handle the problem

[3] This paper is based on the Master Thesis[Glo97].

of unrealistic parallelism in the original active-logic [NKPM97]. In this logic, during a given step i the agent applies its inference rules only to these beliefs of step $i-1$ which are in its short-term memory (denote "focus") at $i-1$. The main intuition behind our formalism is that an agent doesn't draw conclusions based on all of its beliefs. In each time period, only some of its beliefs are in focus and it performs inferences only on these beliefs.

We note that we assume that the focus can consist of very long formulas. In particular, in this paper we concentrate only on the issue of the number of formulas in the focus and not on its size. Restricting the overall length of the formulas in the focus is left for future work.

It is also important to note that our language is a first order modal language and not a second order one, and thus its expression is limited: we are unable to refer to a certain belief at a given reasoning step. For example, it is impossible to express in our language a sentence such as: "There exists a belief in the focus at time-step t".

3 A modal active-logic for reasoning *in* time

This modal logic is based on Montague's semantics [Mon70] which uses structures referred to in the literature as *neighborhood structures* [Che80]. Montague gives a possible world semantic to epistemic logic where, unlike in the Kripke model that suffer omniscience in the form of the K_n axiom $(K\phi \wedge K(\phi \rightarrow \psi) \rightarrow K\psi)$, knowledge is defined as a relation between a world and a set of sets of worlds. An *intension* of a formula ϕ denoted by $\|\phi\|$ is the set of worlds where ϕ is satisfied.

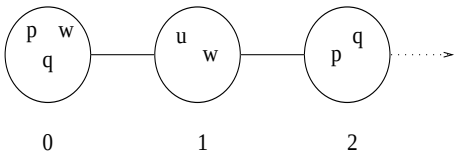


Fig. 1. An example of a time-line. In time-point 0: p, q and w are true, and in time-point 1: u and w are true, etc.

We prefer to use *time-lines* instead of possible worlds, since this gives us a way to naturally incorporate time into our framework. L denotes the set of discrete time-lines [TSSK91]. We consider restricted time-lines to be finite from one side and infinite from the other (i.e., rays). At every time-point in each time-line, some propositions are true and the rest are false (see figure 1).

We use two modal operators modeling belief and focus. The agent is provided with a set of beliefs at time-step 0. This set will be called the initial set of beliefs. The agent is limited – incapable of being aware of all its beliefs simultaneously. Hence, the agent is assisted by the "focus" in order to focus on a limited set of explicit beliefs which it uses in its reasoning process. Only through the focus the

agent draws new beliefs. These new beliefs are entered into the set of beliefs of the next time-point. For example, if the beliefs ψ and $\psi \rightarrow \phi$ are in the focus at time-step t, then the agent believes ϕ at the next time-step. We note that it is possible that the conclusion ϕ is already a belief of the agent. However, if the agent believes only $\phi \rightarrow \psi$ and ϕ but not ψ, it would not believe ψ until $\phi \rightarrow \psi$ and ϕ are brought to focus simultaneously.

3.1 Syntax and semantics

The language $\mathcal{G}$ Formally, we assume that there is a set P of atomic propositions. Let $N = \{0, 1, 2, ...\}$ denote the set of time constants and $\mathcal{V}$ denote the set of time-point variables. We define a set of time terms, T, to be the smallest set containing elements over $N \cup \mathcal{V}$, closed under + sign. We define $P \times T$ to be the set of primitive time propositions. In addition, we use two modal operators B and F. The operator B denotes belief, and the operator F denotes focus of attention.

The language $\mathcal{G}$ is the smallest set of formulas satisfying the following properties[4]:

1. If $\tau_1, \tau_2 \in T$ then $\tau_1 < \tau_2 \in \mathcal{G}$.
2. If $\psi \in P \times T$ then $\psi \in \mathcal{G}$.
3. If $\psi \in \mathcal{G}$ then $\neg\psi \in \mathcal{G}$.
4. If $\psi, \phi \in \mathcal{G}$ then $\psi \wedge \phi, \psi \vee \phi, \psi \rightarrow \phi \in \mathcal{G}$.
5. If $\psi \in \mathcal{G}$ and $\tau \in T$ then $B_\tau\psi \in \mathcal{G}$ and $F_\tau\psi \in \mathcal{G}$.
6. If $\psi \in \mathcal{G}$ and $\tau \in \mathcal{V}$ then $\exists\tau\psi \in \mathcal{G}$ and $\forall\tau\psi \in \mathcal{G}$.

For example, in this language one can express belief formulas such as $B_{\tau_1} p(\tau_2)$ which means "at time τ_1 the agent believes that p is true at time τ_2". It is possible to express nested beliefs formulas such as $B_{\tau_1}(B_{\tau_1+2} p(\tau_2) \vee B_{\tau_1+2} q(\tau_3))$ which means "at time τ_1 the agent believes that two time-points later it will believe $p(\tau_2)$ or it will believe $q(\tau_3)$". Similarly, $F_{\tau_1} p(\tau_2)$ means that "at time τ_1, the belief $p(\tau_2)$ is in the focus". Intuitively concluding new beliefs will be done only on the sentences in F.

This language can easily be extended to include multiple agents, by the use of an additional parameter i, so that $B^i_\tau \alpha$ denotes "at time τ agent i believes in α", where α may include beliefs of other agents.

Modal active-logic structures Let $\mathcal{N}$ to be the set of natural numbers, namely, $\{0, 1, 2, ...\}$.

A structure in the modal active-logic is: $M = \langle L, <, +, u, \pi, \mathcal{B}, \mathcal{F}\rangle$ where

- L is the set of time-lines.
- $<$ is a two-place predicate denoting the usual (strict) ordering relation on $\mathcal{N}$.

[4] For simplicity, we allow only time variables and time terms. It is easy to extend our language to full first order logic with time, belief and focus of attention.

- $+$ is the two-place function denoting the operation 'addition'.
- $u : T \longrightarrow \mathcal{N}$ is a standard interpretation ration of the time terms of the language with respect to the set of natural numbers[5].
- $\pi : P \times \mathcal{N} \times L \longrightarrow \{true, false\}$ is a truth assignment to the atomic formula $p \in P$ for each time-line $l, l \in L$ at the time-point integer $t \in \mathcal{N}$. Thus π defines the intension of the atomic formulas of our language.
- $\mathcal{B} : L \times \mathcal{N} \longrightarrow 2^{2^L}$ is a belief accessibility relation, defined for each time-line and time-point pair (l, t), where $l \in L, t \in \mathcal{N}$.
- $\mathcal{F} : L \times \mathcal{N} \longrightarrow 2^{2^L}$, is a focus accessibility relation, defined for each time-line and time-point pair (l, t), where $l \in L$, $t \in \mathcal{N}$.

We denote by $\mathcal{M}$, the set of all the modal active-logic structures. We will use $\mathcal{B}_t(l)$, $\mathcal{F}_t(l)$ to denote $\mathcal{B}(l, t)$ and $\mathcal{F}(l, t)$ which are the sets of sets of time-lines associated with l at time t through the $\mathcal{B}$ and $\mathcal{F}$ relations, respectively.

We impose restrictions on the accessibility relations of the structures to reflect the step-like reasoning behavior between successive time instances (see section 3.4). We further characterize the modal active-logic by a sound and complete set of axioms and inference rules (see section 3.3).

Satisfiability We formally define satisfiability, $\models$, for the structure $M = \langle L, \mathcal{R}_A, \mathcal{V}, v, \pi, \mathcal{B}, \mathcal{F} \rangle$. Given an extension u which is based on v and a time-line $l \in L$, we denote by $M, l \models_u$ the satisfiability of l of M with respect to u. An intension of a formula ϕ in a structure M, denoted by $\|\phi\|_M$, is $\{l \mid l \in L,\ M, l \models_u \phi\}$. When M is clear from the context, we will omit it and write $\|\phi\|$ (see for example figure 2).

Given a structure M, time-line $l \in L$, and u as described above, if $p \in P$, $\tau_1, \tau_2, \tau \in T$, $\tau' \in \mathcal{V}$, $\phi, \psi \in \mathcal{G}$ then:

1. $M, l \models_u (\tau_1 < \tau_2)$ iff $u(\tau_1) < u(\tau_2)$.
   This defines satisfiability of the relation formulas of our language.
2. $M, l \models_u p(\tau)$ iff $\pi(p(u(\tau)), l) = true$.
   This defines satisfiability of the atomic formulas of our language.
3. $M, l \models_u \neg\phi$ iff $M, l \not\models_u \phi$.
   This defines satisfiability of negated formulas.
4. $M, l \models_u (\phi \wedge \psi)$ iff $M, l \models_u \phi$ and $M, l \models_u \psi$.
   This defines satisfiability of formulas with the $\wedge$ connective.
5. $M, l \models_u B_\tau \phi$ iff $\|\phi\|_M \in \mathcal{B}_{u(\tau)}(l)$.
   This defines the satisfiability of the belief formulas.
6. $M, l \models_u F_\tau \phi$ iff $\|\phi\|_M \in \mathcal{F}_{u(\tau)}(l)$.
   This defines the satisfiability of the focus formulas.
7. $M, l \models_u \forall\tau'\phi$ iff $M, l \models_{u'} \phi$ for every u' which agrees with u everywhere except possibly on τ.
   This defines the satisfiability of the quantified formula.

The satisfiability of $\vee$, $\rightarrow$ and $\exists$ is defined accordingly.

[5] For example, $u(\tau + 5) = u(\tau) + 5$.

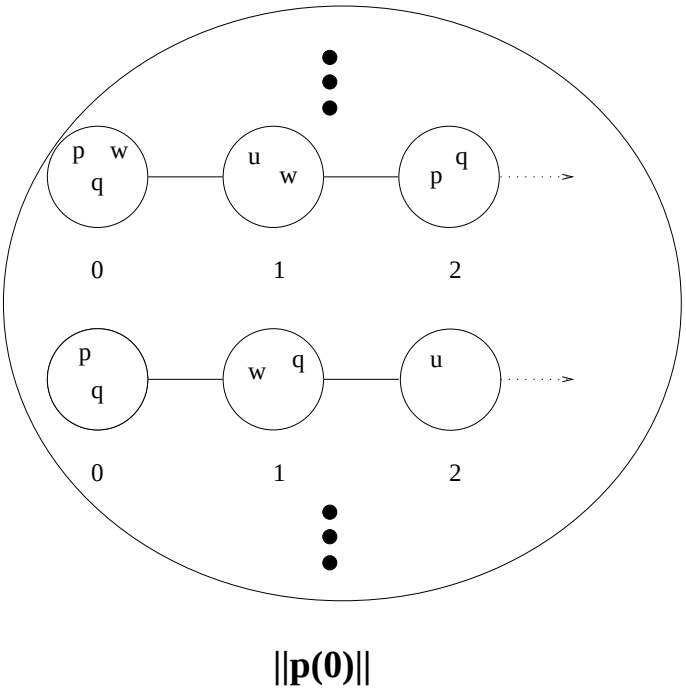


Fig. 2. In this example, we see two of the time-lines of the set $\|p(0)\|$ in which the sentence p is true at time-point 0.

3.2 The process of exposing implicit beliefs

The agent brings some of its beliefs into the focus and reasons with them. At each step it needs to decide which beliefs to bring to the focus. Deciding about the focus can affect the beliefs which the agent will acquire in time, since the agent makes conclusions based only on the formulas in its focus. For example, if the focus does not change over time, then his belief set will not grow over time.

The process presented below provides the agent with the following capability: for each finite subset of beliefs, there exists a finite amount of time (unbounded) within which the agent becomes aware of all the implicit beliefs of this subset. For example, let us examine the finite subset of the initial set of beliefs $\Gamma = \{\psi_1, ..., \psi_n\}$. The agent will deduce all the implicit beliefs of this subset in a finite amount time. Obviously, deducing all the possible consequences of Γ is equivalent to finding all the possible consequences of $\psi_1 \wedge \ldots \wedge \psi_n$.

We use the following definition in the requirements described below.

Definition 1. *Let ψ and ϕ be sentences of $\mathcal{G}$. We say that ψ and ϕ are* **different** *with respect to model M iff $\|\psi\|_M \neq \|\phi\|_M$.*

At the following requirements the focus can include at most two beliefs at each time-step. This restriction is arbitrary, and can be modified to any finite number of sentences.

The requirements on the focus are: (1) Each belief of the initial set of beliefs will be in the focus at a certain time-step, (2) If two beliefs ψ_1, ψ_2 are different, and are in the focus at different time-steps, t_1 and t_2 respectively, then there exists a time-step t that is later than t_1 and t_2 such that both ψ_1 and ψ_2 are in the focus at time-step t and (3) An additional reasoning ability of the agent is that if the agent is aware of a certain belief, ψ, at the time-step t, i.e., ψ is in the focus at time-step t, then the agent believes in all the possible consequences of ψ at the successive time-step $t+1$, i.e., all the possible consequences are in the belief set at time-step $t+1$. Then, if the agent has two different beliefs in the focus at the same time-step t, then the agent adds their conjunction to its beliefs set and its focus set at the next time-step.

To complete the discussion we describe the process of finding all the conclusions from the set $\{\psi_1, ..., \psi_n\}$, where the set $\{\psi_1, ..., \psi_n\}$ is a subset of the initial set of beliefs. According to the first requirement, there exist time-steps $t_1, .., t_n$ (not necessarily different), such that at time-step t_i, belief ψ_i is in the focus ($1 \leq i \leq n$). According to the second requirement and the agent behavior, it is obvious that there exists a time-step t such that $\psi_1 \wedge \ldots \wedge \psi_n$ is in the focus. According to the third requirement, all the consequences of $\psi_1 \wedge \ldots \wedge \psi_n$ become beliefs of the agent at time-step $t+1$. Thus, there exists a time-step in which the agent explicitly believes in all the logical consequences of the set of explicit beliefs $\{\psi_1, ..., \psi_n\}$ which is a finite subset of the initial set of beliefs. A detailed proof appears in the full version of the paper.

3.3 Axiomatization

All axioms and inference rules of the predicate calculus, and the structure natural numbers $\mathcal{R}_A$ [End72] are axioms of our system.

Axiom 1. $B_0 true$.
This means that initially, the agent believes in all the tautologies.

Axiom 2. $\neg B_\tau false$.
This means that the belief operator is consistent at every time point.

Axiom 3. $B_\tau \phi \rightarrow \neg B_\tau(\neg\phi)$.
This ensure that the agent will not have conflicting beliefs in any time period.

Axiom 4. $B_\tau \phi \rightarrow B_{\tau+1}\phi$.
This avoid the forgetting of beliefs.

Axiom 5. $F_\tau \phi \rightarrow B_\tau \phi$.
The focus operator allows focusing on a part of its beliefs in order to conclude from them additional information.

Axiom 6. $B_0 \phi \rightarrow \exists\tau F_\tau \phi$.
This means that every formula the agent initially believes in will enter the focus at a certain time.

Axiom 7. $F_\tau(\phi \wedge \psi) \rightarrow B_{\tau+1}\phi$.

In effect the axiom says that any consequence of a belief in focus should be in implicit belief at the next time. This appears to be an indicated consequence of taking a semantic approach to belief and awareness.

Axiom 8. $\forall\tau_1, \tau_2, F_{\tau_1}\phi \wedge F_{\tau_2}\psi \rightarrow \exists\tau(\tau_1, \tau_2 \leq \tau \wedge F_\tau\phi \wedge F_\tau\psi)$.

We need this axiom to insure that every belief or conclusion that the agent has at any given time-point in the focus will not disappear, but rather will continue to influence its reasoning process.

Axiom 9. $F_\tau\phi \wedge F_\tau\psi \rightarrow F_{\tau+1}(\phi \wedge \psi)$

This means that the pairwise conjunctions of beliefs in the previous focus enter the current step of focus.

We add the following rules:

Rule 1. If $\vdash \phi \leftrightarrow \psi$ then $\vdash B_\tau\phi \leftrightarrow B_\tau\psi$ and $\vdash F_\tau\phi \leftrightarrow F_\tau\psi$.

This rule is inherited by the intensional approach. The agent cannot distinguish between logically equivalent formulas.

Rule 2. If $\nvdash \psi \leftrightarrow \phi$ and $\nvdash \phi \leftrightarrow \chi$ and $\nvdash \psi \leftrightarrow \chi$ then $\vdash F_\tau\phi \wedge F_\tau\psi \rightarrow \neg F_\tau\chi$.

This rule states that the number of different formulas in the focus at a given time-point is limited to two. It is easy to change this axiom to limit the number of different formulas that are in the focus to any given number.

Remark 1. This rule is of course more complicated than the usual logical rules, but there seems to be no easy replacement for it. Still there is a sound possibility of using this rule at least in the finite case (thanks to the reviewer for pointing this out).

3.4 Restriction on the structures

We impose the following restrictions on our structures . These restrictions assure soundness and completeness of the axioms and inference rules described in the previous section.

Restriction 1. For every $l \in L$, $L \in \mathcal{B}_0(l)$.

Note that $\|true\| = L$. This restriction reflects the characterization derived from axiom 1.

Restriction 2. For every $l \in L$, and $t \in \mathcal{N}, \emptyset \notin \mathcal{B}_t(l)$.

This restriction reflects the characterization derived from axiom 2.

Restriction 3. For every $l \in L$, and $s \in \mathcal{B}_0(l)$, there exists a sentence $\phi \in \mathcal{G}$ such that $\|\phi\| = s$. For every $l \in L$, $t \in \mathcal{N}$, $t > 0$, if $s \in \mathcal{B}_t(l)$, then there exists $\phi \in \mathcal{G}$ such that $\|\phi\| = s$ or there exists $t' \in \mathcal{N}$, $t' < t$, $s' \in \mathcal{F}_{t'}(l)$, and $s \supset s'$. For every $l \in L$, $t \in \mathcal{N}$, if $s \in \mathcal{F}_t(l)$, then there exists $\phi \in \mathcal{G}$ such that $\|\phi\| = s$.

This means that every initial belief is expressed by a sentence and every other belief is either expressed by a sentence or is a super set of beliefs that were brought to the focus. In addition, every belief in the focus is expressed by a sentence.

Restriction 4. For every $l \in L$, $t \in \mathcal{N}$, and for every $s_1, s_2 \in \mathcal{B}_t(l)$, if $s_1 \neq s_2$ then $s_1 \cap s_2 \neq \emptyset$.

This restriction reflects the characterization derived from axiom 3.

Restriction 5. For every $l \in L$, and $t \in \mathcal{N}$, $\mathcal{B}_t(l) \subseteq \mathcal{B}_{t+1}(l)$.

This restriction reflects the characterization derived from axiom 4.

Restriction 6. For every $l \in L$, and $t \in \mathcal{N}, \mathcal{F}_t(l) \subseteq \mathcal{B}_t(l)$.

This restriction reflects the characterization derived from axiom 5.

Restriction 7. For every $l \in L$, and $s \in \mathcal{B}_0(l)$, exists $t \in \mathcal{N}$ such that $s \in \mathcal{F}_t(l)$.

This restriction reflects the characterization derived from axiom 6.

Restriction 8. For every $l \in L$, and $t \in \mathcal{N}$, if $s \in \mathcal{F}_t(l)$, and $s' \supseteq s$ then $s' \in \mathcal{B}_{t+1}(l)$.

This restriction reflects the characterization derived from axiom 7 (since supersets correspond to detachments in active-logic).

Restriction 9. For every $l \in L$, and $t_1, t_2 \in \mathcal{N}$, such that $t_1 < t_2$, if $s_1 \in \mathcal{F}_{t_1}(l)$, $s_2 \in \mathcal{F}_{t_2}(l)$ and $s_1 \neq s_2$, then $s_1, s_2 \in \mathcal{F}_t(l)$ for some $t \geq t_2$.

This restriction reflects the characterization derived from axiom 8.

Restriction 10. For every $l \in L$, and $t \in \mathcal{N}$, if $s_1 \in \mathcal{F}_t(l)$, $s_2 \in \mathcal{F}_t(l)$, and $s_1 \neq s_2$, then $s_1 \cap s_2 \in \mathcal{F}_{t+1}(l)$.

This restriction reflects the characterization derived from axiom 9.

Restriction 11. For every $l \in L$, and $t \in \mathcal{N}$, $|\mathcal{F}_t(l)| \leq 2$.

This restriction reflects the characterization derived from rule 2.

Restriction 12. For any sequence sets of atomic formulas SE[6] , sequence of sets of sets of time-lines SB [7] and sequence of sets of sets of sets of time-lines SF [8] which

[6] A *sequence of events* is a sequence of sets of atomic formulas, i.e., $< P_0, P_1, P_2, ... >$ where for all $t \geq 0$, $P_t \subseteq P$. The sequence of events of a time-line $l \in L$ is $< P_0, P_1, ... >$ where for all $t \geq 0$, $P_t = \{p \mid \pi(p, t, l) = true\}$.

[7] A *sequence of beliefs* is a sequence of sets of sets of time-lines, i.e., $< \mathcal{L}_0, \mathcal{L}_1, .. >$ such that for all $t \geq 0$, $\mathcal{L}_t \subseteq 2^L$. The sequence of beliefs of a time-line $l \in L$ is $< \mathcal{B}_0(l), \mathcal{B}_1(l), ... >$.

[8] A *sequence of a focus of attention* is a sequence of sets of sets of time-lines, i.e., $< \mathcal{L}_1, \mathcal{L}_2, .. >$ such that for all $t \geq 0$, $\mathcal{L}_t \subseteq 2^L$. The sequence of the focus of attention of a time-line $l \in L$ is $< \mathcal{F}_0(l), \mathcal{F}_1(l), ... >$.

satisfy restrictions 1-11 above, there is a time-line $l \in L$ such that SE is its sequence of events, SB is its sequence of beliefs and SF is its sequence of focus of attention.

This restriction requires the set L of time-lines to be large enough.

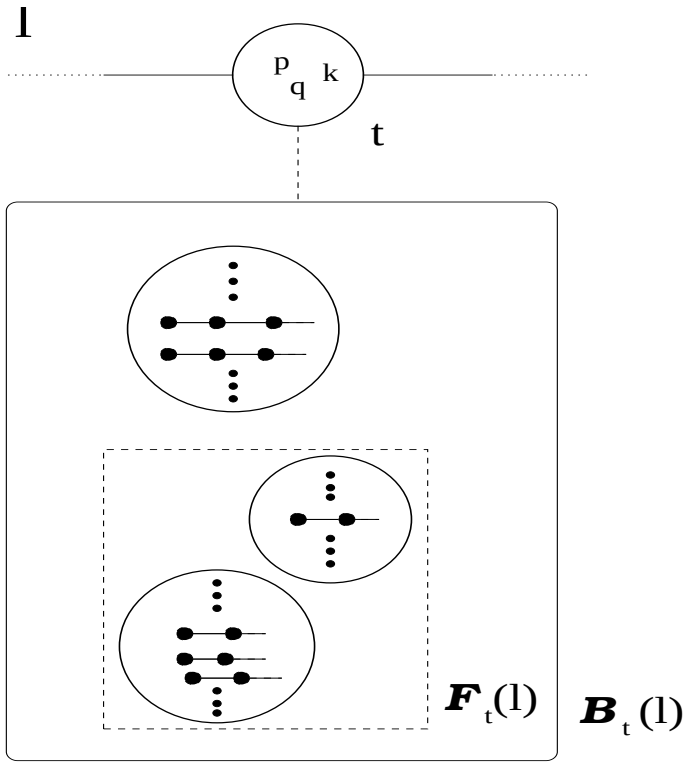


Fig. 3. In this time-line example, at time-point t the atoms p, q, k are true. There are also three different beliefs which are true at time t, while two of them are in the focus set.

In Figure 3 we present an example of a time-point of a time-line in our model. The Soundness and completeness proof for the modal active-logic appear in the full version of the paper.

4 Conclusions and Perspectives

In this paper, we have presented a modal active-logic which accounts for some aspects of the passage of time while the reasoning occurs. To avoid unrealistic parallelism of the agent's reasoning, we introduced the notion of a focus of attention and discussed possible properties of such a system. We characterized the

system by providing a sound and complete axiomatization. In the current model the number of formulas in the focus is restricted, while their length is not. Modeling an agent in which the length of the formulas in its focus is restricted is left for future research. This restriction provokes the question of which beliefs the agent keeps from one step to the next. In particular, we can consider an agent with a short term memory, meaning that it forgets beliefs after some period of time. It will also be important to find a framework in which an agent may believe in only a partial set of tautologies in its language and in which proving tautologies takes time. Another issue to consider is agents able to resolve inconsistencies in their beliefs.

Acknowledgments: We are grateful to Alexander Bochman for instructive comments on this paper.

References

[Ale04] Logan B. Whitsey M. Alechina, N. A complete and decidable logic for resource-bounded agent. *AAMAS 04, Columbia Universty, New-York*, 2004.

[All84] J. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

[Che80] B. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

[EDP90] J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2(1):75–98, 1990.

[End72] Herbert B. Enderton. *A mathmatical introduction to logic*. Academic Press, Inc. Harcourt Brace Jovanovich Publishers, 1972.

[FH88] R. Fagin and J. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39–76, 1988.

[FHV95] R. Fagin, J. Halpern, and M. Y. Vardi. A nonstandard approach to the logical omniscience problem. *Artificial Intelligence*, 79(2):203–240, 1995.

[GFV05] D. Gabbay, M. Fisher, and L. Vila. *Temporal Reasoning in Artifical Intelligence*. Foundations of Artifical Intelligence Series, Volume 1, 2005.

[Glo97] A. Globerman(Botman). *A Modal Active-Logic with Focus of Attention for Reasoning in Time*. Bar Illan University, Isral, 1997.

[KL88] S. Kraus and D. Lehmann. Knowledge, belief and time. *Theoretical Computer Science*, 58:155–174, 1988.

[Kon83] K. Konolige. A deductive model of belief. In *Proceedings of the 8th Int'l Joint Conference on Artificial Intelligence*, pages 377–381, Karlsruhe, West Germany, 1983.

[Lev84] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202, Austin, TX, 1984. AAAI.

[LL00] J. Levesque and G. Lakemeyer. *The logic of knowledge bases*. Cmbridge (MA), London, The MIT Press, 2000.

[McD82] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155, 1982.

[Mon70] R. Montague. Universal grammar. *Theoria*, 36:373–398, 1970.

[Nir94] Madhura Nirkhe. *Time Situated Reasoning within Tight Deadlines and Realistic Space and Computation Bounds*. PhD thesis, University of Maryland, 1994.

[NKPM97] M. Nirkhe, S. Kraus, D. Perlis, and M. Miller. How to (plan to) meet a deadline between *now* and *then*. *Journal of Logic and Computation*, 7(1):109–156, 1997.

[TSSK91] Becky Thomas, Yoav Shoham, Anton Schwartz, and Sarit Kraus. Preliminary thoughts on an agent description language. *International Journal of Intelligent Systems*, 6(5):497–508, August 1991.

[Var86] M. Vardi. On epistemic logic and logical omniscience. In J. Halpern, editor, *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 293–305, Monterey, CA, 1986. Morgan Kaufmann.

Impossible States at Work: Logical Omniscience, Partial Beliefs and Rational Choice

Mikaël Cozic*†

Abstract

Logical omniscience is a never-ending problem in epistemic logic, the main model of *full* beliefs. It is seldom noticed that probabilistic models of *partial* beliefs face the same problem. As far as choice models are built on such doxastic models, they necessarily inherit the problem as well. Following some philosophical ([Hac67]) and decision-theoretic ([Lip99]) contributions, we advocate the use of non-standard or impossible states to tackle this issue. First, we extend the non-standard structures to the probabilistic case ; an axiom system is devised, that is proved to be complete with respect to non-standard probabilistic structures. Second, we show how to substitute weakened doxastic models for the idealized ones in choice models, and discuss the questions raised by this "unidealization".

Keywords : logical omniscience, epistemic logic, probabilistic logic, bounded rationality.

Introduction

Let's imagine an agent that could solve any stochastic decision process, whatever the number of periods, states and alternatives may be ; that could find a Nash Equilibrium in any finite game, whatever the number of players and strategies may be ; more generally, that would have a perfect mathematical knowledge and, still more generally, which would know all the logical consequences of his or her beliefs. By definition, this agent would be described as *logically omniscient*.

For sure, logical omniscience is an highly unrealistic hypothesis from the psychological point of view. Yet, this is the cognitive situation of agents in the main current doxastic models, *i.e.* models of beliefs. The issue has been raised a long time ago in epistemic logic ([Hin75], see the recent survey in [FHMV95]), which is the classical model of *full beliefs*. In particular, it has been recognized that logical omniscience is one of the most uneliminable cognitive idealizations, because it is an immediate consequence of the core principle of the modelling : the representation of beliefs by a space of possible states.

What is the relevance for rational choice theory ? A standard decision model has three fundamental building blocks :

1. a model of beliefs, or doxastic model ;
2. a model of desires, or axiological model, and
3. a criterion of choice, which, given beliefs and desires, selects the "appropriate" actions

In choice under uncertainty, the classical model assumes that the doxastic model is a probability distribution on a state space, the axiological model a utility function on a set of consequences and the criterion is the maximization of expected utility. In this case, the doxastic model is a model of *partial beliefs*. But there are choice models which are built on a model of full beliefs: this is the case of models like maximax or minimax ([LR85], chap.13) where one assumes that the agent takes into account the subset of possible states that is compatible with his or her beliefs.

The point is that, in both cases, *the choice model inherits the cognitive idealizations of the doxastic model*. Consequently, the choice model is cognitively *at least as unrealistic as* the doxastic model upon

*ENS Ulm, Department of Philosophy and Paris IV Sorbonne. 45, rue d'Ulm, F-75005 Paris. Mikael.Cozic@paris4.sorbonne.fr.

†I would like to thank for their helpful comments D.Andler, A. Barberousse, A. Barton, D.Bonnay, I. Drouet, P. Egré, B. Kooi, Ph. Mongin, C. Paternotte, B.Walliser and three anonymous referees. I am also indebted to the seminar participants of "Formal Philosophy" and "Philosophy of Probability" (both at IHPST) ; and colloquium participants of ECCE1 (Gif-sur-Yvette, september 2004), the First Congress of the Society of Philosophy of Sciences (Paris, january 2005) and the First Paris-Amsterdam Meeting of Young Researchers (Amsterdam, june 2005).

which it is based. Indeed, a choice model is strictly more unrealistic than its doxastic model since it assumes furthermore the axiological model and the implementation of the choice criterion. Hence, one of the main sources of cognitive idealization in choice models is the logical omniscience of their doxastic model ; the weakening of logical omniscience in a decision-theoretic context is therefore one of the main ways to build more realistic choice models, *i.e.* to achieve bounded rationality.

Surprisingly, whereas there has been extensive work on logical omniscience in epistemic logic, there has been very few attempts to investigate the extension of the putative solutions to the probabilistic representation of beliefs (*probabilistic case*) and to models of decision making (*decision-theoretic case*)[1].

The aim of this paper is to make some progress in filling this gap. Our method is the following one: given that a huge number of (putative) solutions to logical omniscience have been proposed in epistemic logic, we won't start from scratch, but we will consider extensions of the main current solutions. Our main claim is that the solution that we will call the "non-standard structures" constitute the best candidate to this extension.

The remainder of the paper proceeds as follows. In section 1, the problem of logical omniscience and its most popular solutions are briefly recalled. Then, it shall be argued that, among these solutions, non-standard structures are the best basis for an extension to probabilistic and decision-theoretic cases. Section 2 is devoted to the probabilistic case and states our main result: an axiomatization for non-standard explicit probabilistic structures. In section 3, we discuss the extension to the decision-theoretic case.

1 Logical omniscience in epistemic logic

1.1 Epistemic logic

Problems and propositions related to logical omniscience are best expressed in a logical framework, usually called "epistemic logic" (see [FHMV95] for an extensive technical survey and [Sta91], reprinted in [Sta98] for an illuminating philosophical discussion), which is nothing but a particular interpretation of modal logic. Here is a brief review of the classical model: Kripke structures.

First, we have to define the *language* of propositional epistemic logic. The only difference with the language of propositional logic is that this language contains a doxastic operator B : $B\phi$ is intended to mean "the agent believes that ϕ ".

Definition 1
The set of ***formulas*** *of an epistemic propositional language $\mathcal{L}B(At)$ based on a set At of propositional variables $Form(\mathcal{L}B(At))$, is defined by*

$$\phi ::= p|\neg\phi|\phi \vee \psi|\phi \wedge \psi|B\phi$$ [2]

The interpretation of the formulas is given by the famous Kripke structures :

Definition 2
Let $\mathcal{L}B(At)$ an epistemic propositional language ; a ***Kripke structure*** *for $\mathcal{L}B(At)$ is a 3-tuple $\mathcal{M} = (S, \pi, R)$ where*

(i) S is a state space,

(ii) $\pi : At \times S \rightarrow \{0, 1\}$ is a valuation

(iii) $R \subseteq S \times S$ is an accessibility relation

Intuitively, the accessibility relation associates to every state the states that the agent considers possible given his or her beliefs. π associates to every atomic formula, in every state, a truth value ; it is extended in a canonical way to every formula by the satisfaction relation.

Definition 3
The ***satisfaction relation****, labelled $\models$, extends π to every formula of the language according to the following conditions :*

[1] One important exception is [Lip99].

[2] This formulation (the so-called Backus-Naur Form) means, for instance, that the propositional variables are formulas, that if ψ is a formula, $\neg\psi$ too, and so on.

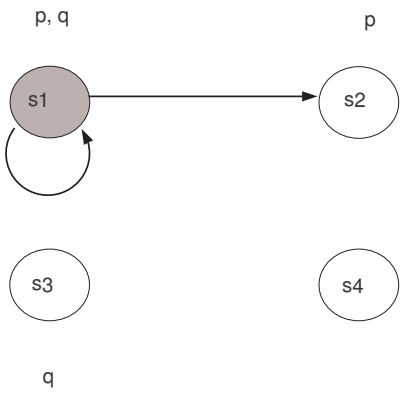


Figure 1: Kripke structures

(i) $\mathcal{M}, s \vDash p$ *iff* $\pi(p, s) = 1$

(ii) $\mathcal{M}, s \vDash \phi \wedge \psi$ *iff* $\mathcal{M}, s \vDash \phi$ *and* $\mathcal{M}, s \vDash \psi$

(iii) $\mathcal{M}, s \vDash \phi \vee \psi$ *iff* $\mathcal{M}, s \vDash \phi$ *or* $\mathcal{M}, s \vDash \psi$

(iv) $\mathcal{M}, s \vDash \neg\phi$ *iff* $\mathcal{M}, s \nvDash \phi$

(v) $\mathcal{M}, s \vDash B\phi$ *iff* $\forall s'$ *s.t.* sRs', $\mathcal{M}, s' \vDash \phi$

The specific doxastic condition contains what might be called the **possible-state analysis of belief**. It means that an agent believes that ϕ if, in all the states that (according to him or her) could be the actual state, ϕ is true : *to believe something is to exclude that it could be false*. Conversely, an agent doesn't believe ϕ if, in some of the states that could be the actual state, ϕ is false : *not to believe is to consider that it could be false*. This principle will be significant in the discussions below.

Example 1
$S = \{s_1, s_2, s_3, s_4\}$; p ("it's sunny") is true in s_1 and s_2, q ("it's windy") in s_1 and s_4. Suppose that s_1 is the actual state and that in this state the agent believes that p is true but does not know if q is true. Figure 1 represents this situation, omitting the accessibility relation in the non-actual states.

Definition 4
Let $\mathcal{M}$ be a Kripke structure ; in $\mathcal{M}$, the set of states where ϕ is true, or the ***proposition*** *expressed by ϕ, or the* ***informational content*** *of ϕ, is noted* $[[\varphi]]_{\mathcal{M}} = \{s : \mathcal{M}, s \vDash \varphi\}$.

To formulate logical omniscience, we need lastly to define the following semantical relations between formulas.

Definition 5
ϕ $\mathcal{M}$***-implies*** ψ *if* $[[\phi]]_{\mathcal{M}} \subseteq [[\psi]]_{\mathcal{M}}$. ϕ *and* ψ *are* $\mathcal{M}$***-equivalent*** *if* $[[\phi]]_{\mathcal{M}} = [[\psi]]_{\mathcal{M}}$

There are several forms of logical omniscience (see [FHMV95]) ; the next proposition shows that two of them, deductive monotony and intensionality, hold in Kripke structures :

Proposition 1
Let $\mathcal{M}$ be a Kripke structure and $\phi, \psi \in \mathcal{L}B(At)$;

(i) ***Deductive monotony*** *: if ϕ $\mathcal{M}$-implies ψ, then $B\phi$ $\mathcal{M}$-implies $B\psi$*

(ii) ***Intensionality*** *: if ϕ and ψ are $\mathcal{M}$-equivalent, then $B\phi$ and $B\psi$ are $\mathcal{M}$-equivalent*

Both properties are obvious theorems in the axiom system K, which is sound and complete for Kripke structures :

System K
(PROP) Instances of propositional tautologies
(MP) From ϕ and $\phi \rightarrow \psi$ infer ψ
(K) $B\phi \wedge B(\phi \rightarrow \psi) \rightarrow B\psi$
(RN) From ϕ, infer $B\phi$

1.2 Three solutions to logical omniscience

A huge number of solutions have been proposed to weaken logical omniscience, and arguably no consensus has been reached (see [FHMV95])[3]. We identify three main solutions to logical omniscience, which are our three candidates to an extension to the probabilistic or decision-theoretic case. There is probably some arbitrariness in this selection, but they are among the most used, natural and powerful existing solutions.

1.2.1 Neighborhood structures

The "neighborhood structures", sometimes called "Montague-Scott structures" are our first candidate. The basic idea is to make explicit the *propositions* that the agent believes ; the neighborhood system of an agent at a given state is precisely the set of propositions that the agent believes.

Definition 6
*A **neighborhood structure** is a 3-tuple $\mathcal{M} = (S, \pi, V)$ where*

(i) S is a state space,

(ii) $\pi : At \times S \rightarrow \{0, 1\}$ is a valuation,

*(iii) $V : S \rightarrow \wp(\wp(S))$, called the agent's **neighborhood system**, associates to every state a set of propositions.*

The conditions on the satisfaction relation are the same, except for the doxastic operator :

$$\mathcal{M}, s \vDash B\phi \text{ iff } [[\phi]]_{\mathcal{M}} \in V(s)$$

It's easy to check that deductive monotony is invalidated by neighborhood structures, as shown by the following example.

Example 2
Let's consider the first example and replace the accessibility relation by a neighborhood system ; $V(s_1)$ contains $\{s_1, s_2\}$ but not $\{s_1, s_2, s_3\}$. Then, in s_1, Bp is true but not $B(p \vee q)$. This is represented in Figure 2[4]*.*

As expected, one can regain deductive monotony by closing the neighborhood systems under supersets. Nonetheless, the axiomatization presented below[5] makes clear that the power of neighborhood system is limited: intensionality cannot be weakened.

[3] We have contributed to this industry by defending the use of substructural logics in [Coz06] ; this setting is not tractable enough for the aim of the current paper.

[4] This reccuring example is not chosen for its cognitive realism, but because it makes the comparison of different solutions easy.

[5] The system E is strong and complete with respect to neighborhoods structures.

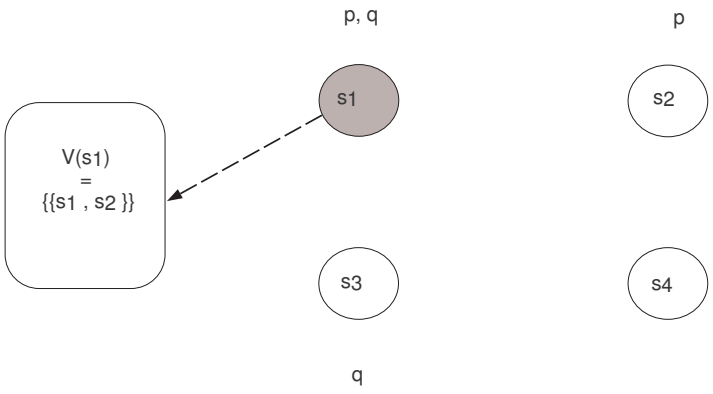


Figure 2: Neighborhood structures

.

System E (Chellas, 1980)
(PROP) Instances of propositional tautologies
(MP) From ϕ and $\phi \rightarrow \psi$ infer ψ
(RE) From $\phi \leftrightarrow \psi$ infer $B\phi \leftrightarrow B\psi$

1.2.2 Awareness structures

The second solution, due to J. Halpern and R. Fagin ([FH88][6]) , are the "awareness structures". The basic idea is to put a *syntactical filter* on the agent's beliefs. The term "awareness" suggests that this can be interpreted as reflecting the agent's awareness state, but other interpretations are conceivable as well.

Definition 7
*An **awareness structure** is a 4-tuple* (S, π, R, A) *where*

(i) S *is a state space,*

(ii) $\pi : At \times S \rightarrow \{0, 1\}$ *is a valuation,*

(iii) $R \subseteq S \times S$ *is an accessibility relation,*

(iv) $A : S \rightarrow Form(\mathcal{L}B(At))$ *is a function which maps every state in a set of formulas ("awareness set").*

The new condition on the satisfaction relation is the following:

$\mathcal{M}, s \vDash B\phi$ iff $\forall s'$ s.t. sRs', $s' \in [[\phi]]_{\mathcal{M}}$ and $\phi \in A(s)$

This new doxastic condition permits to weaken *any form* of logical omniscience ; in particular, our example shows how to model an agent who violates deductive monotony.

Example 3
Let's consider our example and stipulate that $A(s_1) = \{p\}$*. Then it is still the case that* Bp *is true in* s_1*, but not* $B(p \vee q)$*. This is represented in Figure 3.*

If one keeps the basic language $\mathcal{L}B(At)$, one obtains as axiom system a minimal epistemic logic which eliminate any form of logical omniscience:

[6]What we call "awareness structures" is called in the original paper "logic of general awareness".

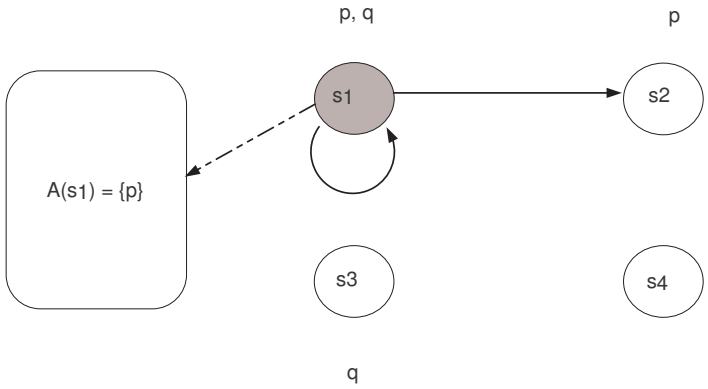


Figure 3: Awareness structures

Minimal Epistemic Logic (FHMV 1995)
(PROP) Instances of propositional tautologies (MP) From ϕ and $\phi \rightarrow \psi$ infer ψ

1.2.3 Non-standard structures

We now switch to our last solution : the non-standard structures, which are sometimes called "Kripke structures with impossible states". Contrary to the two preceding solutions, neither the accessibility relation nor the doxastic condition are modified. What is revised is the underlying state space or, more precisely, the nature of the satisfaction relation in certain states of the state space.

Definition 8
*A **non-standard structure** is a 5-tuple $\mathcal{M} = (S, S', \pi, R, \vDash)$ where*

(i) S is a space of standard states,

(ii) S' is a space of non-standard states,

(iii) $R \subseteq S \cup S' \times S \cup S'$ is an accessibility relation,

(iv) $\pi : Form(\mathcal{L}B(At)) \times S \rightarrow \{0, 1\}$ is a valuation on S

(v) $\vDash$ is a satisfaction relation which is standard on S (recursively defined as usual) but arbitrary on S'

In non-standard structures, there are no *a priori* constraints on the satisfaction relation in non-standard states. For instance, in a non-standard state s', both ϕ and $\neg\phi$ can be false. For every formula ϕ, one might therefore distinguish its *objective informational content* $[[\phi]]_{\mathcal{M}} = \{s \in S : \mathcal{M}, s \vDash \phi\}$ from its *subjective informational content* $[[\phi]]^*_{\mathcal{M}} = \{s \in S^* = S \cup S' : \mathcal{M}, s \vDash \phi\}$. In spite of appearances, this generalization of Kripke structures is arguably natural as soon as one accepts the possible-state analysis of beliefs. Recall that, according to this analysis,

- to believe that ϕ is to exclude that ϕ could be false, and
- not to believe that ψ is not to exclude that ψ could be false.

In consequence, according to the possible-state analysis, to believe that ϕ but not to believe one of its logical consequences ψ is to consider as possible at least one state where ϕ is true but ψ false. By definition, a state of this kind is logically non-standard. Non-standard structures is the most straightforward way to keep the possible-state analysis of beliefs[7].

[7] For a more extensive defense of the solution, see [Hin75] or, more recently, [Bar97].

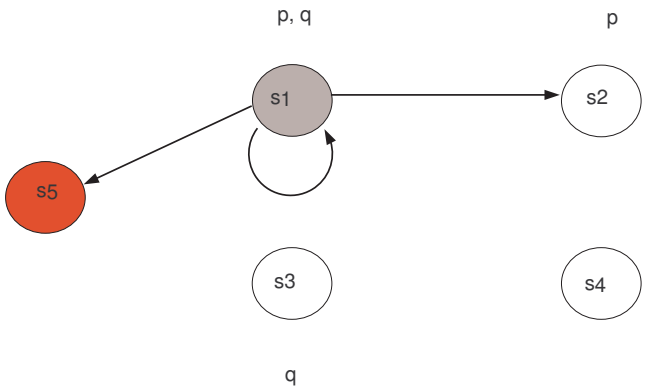


Figure 4: Non-standard structures

Example 4
Let's consider our example but add a non-standard state in $S' = \{s_5\}$; we stipulate that $\mathcal{M}, s_5 \vDash p$, but that $\mathcal{M}, s_5 \nvDash (p \vee q)$. Then in s_1, Bp is true but not $B(p \vee q)$. This is represented in Figure 4.

2 The probabilistic case

Mainstream decision theory is based on doxatic models of *partial beliefs*, not of full beliefs. Hence weakenings of logical omniscience in the framework of doxastic logic does not give directly a way to weaken logical omniscience that is appropriate for decision theorists. The aim of this section is to study the probabilistic extension of doxastic models without logical omniscience.

2.1 Probabilistic counterpart of logical omniscience

First, we have to define the probabilistic counterparts of logical omniscience. In the usual (non-logical) framework, if P is a probability distribution on S[8], then the following property is the counterpart of logical omniscience : if $E \subseteq E'$, then $P(E) \leq P(E')$,

But to be closer to the preceding section, it is better to work with an elementary[9] logical version of the usual probabilistic model :

Definition 9
*Let $\mathcal{L}(At)$ a propositional language ; a **probabilistic structure**[10] for $\mathcal{L}(At)$ is a 3-tuple $\mathcal{M} = (S, \pi, P)$ where*

(i) S is a state space,

(iii) π is a valuation,

(iv) P is a probability distribution on S.

We will say that an agent believes to degree r a formula $\phi \in Form(\mathcal{L}(At))$, symbolized by $CP(\phi) = r$, if $P([[\phi]]_{\mathcal{M}}) = r$[11]. We can state the precise probabilistic counterparts of logical omniscience:

Proposition 2
The following holds in probabilistic structures :

[8] In the paper, to avoid complications that are unnecessary for our purpose, we suppose that S is finite and that P is defined on $\wp(S)$.
[9] "Elementary" because there is no doxastic operator in the object-language.
[10] See [FH91]. For a recent reference on logical formalization of probabilistic reasoning, see [Hal03].
[11] Note that $CP(\phi) = r$ is in the meta-language, not in the object-language.

(i) deductive monotony : if ϕ $\mathcal{M}$-implies ψ, then $CP(\phi) \leq CP(\psi)$.

(ii) intensionality : if ϕ and ψ are $\mathcal{M}$-equivalent, then $CP(\phi) = CP(\psi)$.

One can check that these are indeed the *counterparts* of logical omniscience by looking at the limit case of *certainty*, *i.e.* of maximal degree of belief : (i) if an agent is certain that ϕ and if ϕ $\mathcal{M}$-implies ψ, then the agent is certain that ψ as well ; (ii) if ϕ and ψ are $\mathcal{M}$-equivalent, then an agent is certain that ϕ iff he or she is certain that ψ.

Which of the three solutions to choose for this extension ?

(a) First, we should eliminate neighbordhood structures because their power is limited: intensionality is a too strong idealization. This is especially sensitive in a decision context, where, under the label of "framing effects", it is has been recognized for a long time that logically equivalent formulations of a decision problem could lead to different behaviors.

(b) Second, the extension of awareness structures seems intrisically tricky. Suppose that an agent believes ϕ to degree r_ϕ and ψ to degree r_ψ with ϕ $\mathcal{M}$-implying ψ and $r_\phi > r_\psi$. This is a failure of deductive monotony. Now, in an analogous situation, the way awareness structures proceed in epistemic logic is by "dropping" the formula ψ. Let's apply this method to the probabilistic case: we would say that an agent believes that ϕ to degree r if $P([[\phi]]_\mathcal{M}) = r$ *and* he or she is aware of ϕ. But no one could model a situation like the preceding one: either the agent is aware of ψ and in this case necessarily he or she believes that ψ to a degree $r_\psi \geq r_\phi$; or he or she is not aware of ψ, and it this case he or she has no degree of belief toward ψ. This is not a knock-down argument, but it implies that if one wants to extend awareness structures, one has to make it substantially more sophisticated.

(c) Lastly, the extension of awareness structure is problematic in our perspective, *i.e.* a perspective of decision-theoretic application. To see why, let's notice that a criterion choice like expected utility might be seen as a function whose first argument is a doxastic model and second argument an axiological model. If we would extend the awareness structures, the first value of an expected utility criterion would not be any more a simple probability distribution. Consequently, *we should have to revise our choice criterion*. For sure, nothing precludes such a move, but simplicity recommends another tactic.

We are therefore left with non-standard structures. Non-standard structures do not suffer from the above mentioned troubles : they are as powerful as one can wish, the extension is intrinsically simple and they should permit to keep usual choice criterion when embedded in a choice model. This is our motivation, but now we have to turn to positive arguments[12].

2.2 Non-standard implicit probabilistic structures

To give the basic insights and show the fruitfulness of the proposition, we will continue to work in the elementary setting where no doxastic operators are in the object-language.

Definition 10
Let $\mathcal{L}(At)$ a propositional language ; a ***non-standard implicit probabilistic structure*** *for $\mathcal{L}(At)$ is a 5-tuple $\mathcal{M} = (S, S', \pi, \models, P)$ where*

(i) S is a standard state space,

(ii) S' is a non-standard space,

(iii) $\pi : Form(L(At)) \times S \to \{0, 1\}$ is a valuation on S,

(iv) $\models$ is a satisfaction relation which is standard on S but arbitrary on S'

(v) P is a probability distribution on $S^* = S \cup S'$.

As in the set-theoretic case, one can distinguish the objective informational content of a formula, *ie* the standard states where this formula is true, and the subjective informational content of a formula, *ie* the states where this formula is true.

To obtain the expected benefit, the non-standard probabilistic structures should characterize the agent's doxastic state on the basis of *subjective* informational content : an agent believes a formula ϕ to degree r,

[12] A similar idea has been defended a long time ago by I. Hacking who talks about "personal possibility", by contrast with "logical possibility". We won't develop the point here, but this contribution can be seen as a formalization of Hacking's insights ([Hac67]).

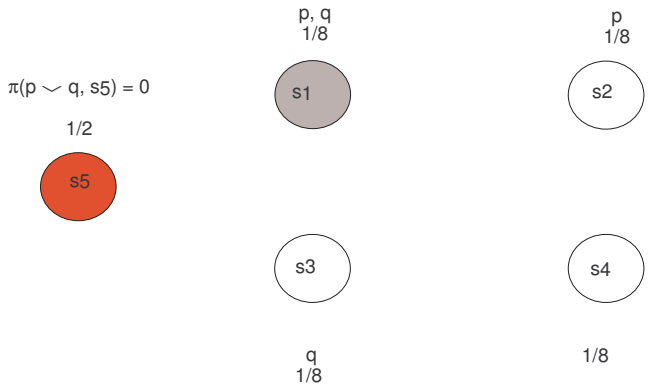


Figure 5: Probabilistic non-standard structures

$CP(\phi) = r$, if $P([[\phi]]^*_{\mathcal{M}}) = r$. It is easy to check that, in this case, logical omniscience can be utterly controlled.

Example 5
Let's take the same space state as in the preceding examples. Suppose that the agent has the following partial beliefs : $CP(p) > CP(p \vee q)$. *This can be modelled in the following way :* $S' = \{s_5\}$, $s_5 \in [[p]]^*_{\mathcal{M}}$ *but* $s_5 \notin [[p \vee q]]^*_{\mathcal{M}}$, $P(s_1) = P(s_2) = P(s_3) = P(s_4)1/8$ *and* $P(s_5) = 1/2$. *This is represented in Figure 5.*

2.3 Special topics : deductive information and additivity

This extension of non-standard structures is admittedly straightforward and simple. It gives immediately the means to weaken logical idealizations. Furthermore, it opens perspectives specific to the probabilistic case; two of them will be briefly mentioned.

Deductive information and learning First, one can model the fact that an agent acquires not only empirical information but *deductive information*; in non-standard structures, this corresponds to the fact that *the agent eliminates non-standard states*.

Let's come back to our generic situation. Suppose that our agent learns that ϕ implies ψ. This means that he or she learns that the states where ϕ is true but ψ false are impossible. This is equivalent to say that he or she learns the event

$$I = S^* - ([[\phi]]^*_{\mathcal{M}} - [[\psi]]^*_{\mathcal{M}})$$

To be satisfying, such a notion of deductive information must respect a requirement of compatibility between revising and logical monotony: if the agent learns that ϕ implies ψ and revise his or her beliefs upon this fact, his or her new probability distribution should conform to logical monotony with respect to ϕ and ψ. One can check that it is the case with the main revising rule, *i.e.* conditionalization.

Proposition 3
If I *is learned following the conditionalization, then deductive monotony is regained,* ie $CP_I(\phi) \leq CP_I(\psi)$.

Example 6
This can be checked in the preceding example : $I = S = \{s_1, s_2, s_3, s_4\}$. *By conditionalization,* $CP_I(p) = 1/2$ *whereas* $CP_I(p \vee q) = 3/4$.

Additivity A second topic is additivity. From a logical point of view, one can define additivity as follows :

Definition 11
$\mathcal{M}$ is (logically) ***additive*** *if, when ϕ and ψ are logically incompatible, $CP(\phi) + CP(\psi) = CP(\phi \vee \psi)$.*

Additivity is of course the core of the probabilistic representation of beliefs, and alternative representations of beliefs depart often from probability on this point. For example, in the Dempster-Shafer theory ([Sha76]), the so-called belief function is superadditive (in our notation, $CP(\phi \vee \psi) \geq CP(\phi) + CP(\psi)$) whereas its dual, the plausibility function, is subadditive ($CP(\phi \vee \psi) \leq CP(\phi) + CP(\psi)$).

A noteworthy aspect of probabilistic non-standard structures is that the freedom of the connectives' behavior in non-standard states permits us to have a very flexible framework with respect to additivity : simple conditions on the connectives imply general properties concerning additivity.

Definition 12
*Let $\mathcal{M} = (S, S', \pi, \vDash, P)$ a non-standard probabilistic structure ; $\mathcal{M}$ is $\vee$-**standard** if for every formulas ϕ, ψ, $[[\phi \vee \psi]]^*_{\mathcal{M}} = [[\phi]]^*_{\mathcal{M}} \cup [[\psi]]^*_{\mathcal{M}}$.*

This means that the disjunction behaves in the usual way in non-standard states ; a trivial consequence of this is that the structure $\mathcal{M}$ is (logically) subadditive.

Proposition 4
If $\mathcal{M}$ is $\vee$-standard, then it is (logically) subadditive.

To be a little bit more general, one can consider the (logical) inclusion-exclusion rule :

$$CP(\phi \vee \psi) = CP(\phi) + CP(\psi) - CP(\phi \wedge \psi)$$

One can define (logical) **submodularity** (resp. supermodularity or convexity) as : $CP(\phi \vee \psi) \leq CP(\phi) + CP(\psi) - CP(\phi \wedge \psi)$ (resp. $CP(\phi \vee \psi) \geq CP(\phi) + CP(\psi) - CP(\phi \wedge \psi)$).

It's clear that to control submodularity, we have to control the conjunction's behavior.

Definition 13
Let $\mathcal{M} = (S, S', \pi, \vDash, P)$ a probabilistic non-standard structure ;

(i) $\mathcal{M}$ is ***negatively*** *$\wedge$-**standard** if for every formulas ϕ, ψ, when $\mathcal{M}, s \nvDash \phi$ or $\mathcal{M}, s \nvDash \psi$, then $\mathcal{M}, s \nvDash \phi \wedge \psi$.*

(ii) $\mathcal{M}$ is ***positively*** *$\wedge$-**standard** if for every formulas ϕ, ψ, when $\mathcal{M}, s \vDash \phi$ and $\mathcal{M}, s \vDash \psi$, then $\mathcal{M}, s \vDash \phi \wedge \psi$.*

Proposition 5
Suppose that $\mathcal{M}$ is $\vee$-standard ;
- if $\mathcal{M}$ is negatively $\wedge$-standard, then submodularity holds.
- if $\mathcal{M}$ is positively $\wedge$-standard, then supermodularity holds.

Proof : see the Appendix.

2.4 Non-standard explicit probabilistic structures

Implicit probabilistic structures are not very expressive ; to have a true analogon of epistemic logic, we have to start from an object-language that contains (partial) doxastic operator.

Following R. Aumann ([Aum99]) and A. Heifetz and Ph. Mongin ([HM01]), we consider the operator L_a[13][14]. The intuitive meaning of $L_a\phi$ is: the agent believes at least to degree a that ϕ. Note that we add the usual symbols $\top$, $\bot$: $\top$ is what the agent recognizes as necessarily true and $\bot$ is what he or she recognizes as necessarily false.

Definition 14
The set of formulas of an explicit probabilistic language $\mathcal{L}L(At)$ based on a set At of propositional variables, $Form(\mathcal{L}L(At))$ is defined by :

[13] Economists are leading contributors to the study of explicit probabilistic structures because they correspond to the so-called *type spaces* used in games of incomplete information, in the same way that Kripke structures (with R as an equivalence relation). See [AH02].

[14] Note that another language is used by J. Halpern in [FH91] or [Hal03].

$$\phi ::= p|\bot|\top|\neg\phi|\phi \vee \psi|L_a\phi$$

where $p \in At$ *and* $a \in [0,1] \subseteq \mathbb{Q}$.

The corresponding structures are an obvious extension of implicit non-standard structures:

Definition 15
A ***non-standard explicit probabilistic structure*** *for* $\mathcal{L}L_a(At)$ *is a 5-tuple* $\mathcal{M} = (S, S'\pi, \vDash, P)$ *where*

(i) $\vDash$ *is a satisfaction relation s.t.*

(a) $\vDash$ *is standard on* S *for all propositional connectives*
(b) $\forall s \in S$, $\mathcal{M}, s \vDash L_a\phi$ *iff* $P(s)([[\phi]]^*_{\mathcal{M}}) \geq a$
(c) $\forall s \in S \cup S'$, $\mathcal{M}, s \vDash \top$ *and* $\mathcal{M}, s \nvDash \bot$

(ii) $P : S^* \to \Delta(S^*)$ *assigns to every state a probability distribution on the state space.*

In [Aum99], R. Aumann has failed to axiomatize (standard) explicit probabilistic structures, but [HM01] have recently devised an axiom system that is (weakly) complete for these structures. In comparison with epistemic logic, one of the problems is that the adaptation of the usual proof method, *i.e.* the method of canonical models, is not trivial. More precisely, in the epistemic logic's case, it is easy to define a canonical accessibility relation on the canonical state space. This is not case in the probabilistic framework, where strong axioms are needed to guarantee that. Fortunately, the non-standard structures permit huge simplifications, and one can devise an axiom system that essentially mimics the Minimal Epistemic Logic above described.

Minimal Probabilistic Logic
(PROP) Instances of propositional tautologies
(MP) From ϕ and $\phi \to \psi$ infer ψ
(A1) $L_0\phi$
(A2) $L_a\top$
(A2+) $\neg L_a\bot$ $(a > 0)$
(A7) $L_a\phi \to L_b\phi$ $(b < a)$

The axioms' notation follows [HM01] to facilitate comparison. Axioms (A2) and (A2+) reflect our semantic for $\top$ and $\bot$: the agent believes to maximal degree what he or she considers as necessarily true and does not believe to any degree what he or she considers as necessarily false. (A1) and (A7) reflect principles specific to the probabilistic case. Note that both bear on a single embedded formula ϕ: there is no doxastic reflection of a logical relation. They express something like a minimal metric of partial beliefs.

If $\vDash_{NSEPS} \phi$ means that ϕ is true in every non-standard explicit probabilistic structure and $\vdash_{MPL} \phi$ that ϕ is provable in the Minimal Probabilistic Logic, then we are ready to state our main result:

Theorem 1 (Soudness and Completeness of MPL)

$$\vDash_{NSEPS} \phi \text{ iff} \vdash_{MPL} \phi$$

Proof : see the Appendix.

3 Insights into the decision-theoretic case

We would like to end this paper by showing how to build choice models without logical omniscience, and which are the challenges raised by such a project.

3.1 Choice models without logical omniscience

The basic method to build a choice model without logical omniscience is to *substitute* one of our non-standard structures to the original doxastic model in the target choice model. We will now show how this could be done.

One might generically see models of choice under uncertainty as based on

- a state space S

- a set A of actions
- a consequence function $\mathfrak{C} : S \times A \to C$ where C is a set of consequences
- a utility function $u : C \to \mathbb{R}$
- a criterion of choice

To complete the choice model, one adds a distribution P on S for models of choice under probabilistic uncertainty, and a set $K \subseteq S$ of states compatible with the agent's beliefs under set-theoretic uncertainty.

To rigourously extend non-standard structures to choice models, one should translate the above described notions in a logical setting. But to give some insights, we will, on the contrary, import non-standard structures in the syntax-free framework of conventional decision theory. Let's have a look at the following, admittedly particular, target situation : an agent knows abstractly the consequence function $\mathfrak{C}$, but, because of limited computational capacities,he or she is not able, at the moment of choice, to perfectly infer from the choice function the consequence of each action at each possible state. One can think about a classic two-state example of insurance application [15]. The consequence function is

$\mathfrak{C}(s_1, x) = w - \pi x$

$\mathfrak{C}(s_2, x) = y + x$,

where x, the choice variable, is the amount of money spent in insurance, s_1 the state without disaster, w the wealth in s_1, s_2 the state with a disaster and y the subsequent wealth, and π the rate of exchange. In this case, a non logically omniscient agent with respect to the consequence function would be such that he or she ignores the value of $\mathfrak{C}$ for some arguments.

A simple way to model this target situation would be the following one. Let's consider *extended states* w, which are composed of a (primitive) state s and a local consequence function $\mathfrak{C}_w : A \to C$: $w = (s, \mathfrak{C}_w)$. The set of extended states is intended to represent the beliefs of the agent, including his or her logically imperfect beliefs. An extended state is standard if its local consequence function is conform to the (true) consequence function: $\mathfrak{C}_w(a) = \mathfrak{C}(a, s)$; if not, it is non-standard.

For instance, a logically imperfect agent could not know what is the consequence of action a in state s, thinking that it is possible that this consequence is c_i (let's say, the true one) or c_j. This situation would be modeled by building (at least) two extended states :

- $w_i = (s, \mathfrak{C}_{w_i})$ where $\mathfrak{C}_{w_i}(a) = c_i$, and
- $w_j = (s, \mathfrak{C}_{w_j})$ where $\mathfrak{C}_{w_j}(a) = c_j$

A perfect logician wouldn't have considered a possible state like w_j. On this basis, one can build choice models without the assumption of logical omniscience:

- in the case of choice under set-theoretic uncertainty, if one takes the maximin criterion, for a belief set $K \subseteq W$, the solution is :

$$Sol(A, S, W, C, \mathfrak{C}, u, K) = \arg\max_{a \in A} \min_{w \in K} u(\mathfrak{C}_w(a))$$

- in the case of choice under probabilistic uncertainty, if one takes the maximization of expected utility criterion, for a probability distribution P on W, the solution is :

$$Sol(A, S, W, C, \mathfrak{C}, u, P) = \arg\max_{a \in A} \sum_{w \in W} P(w).u(\mathfrak{C}_w(a))$$

3.2 Open questions

From the decision theorists point of view, the substitution we have just described is only a first step. Two fundamental questions remains.

(a) First, there is the question of the axiomatization of the new choice models, that is closely linked with the behavioral implications of choice models without logical omniscience. In a recent paper, B. Lipman ([Lip99]) has remarkably tackled this issue, advocating a very similar approach. But the choice model he uses is quite specific (conditional expected utility), and one would like to compare choice models based on non-standard structures with the savagean benchmark.

More precisely, one would like to obtain a *representation theorem* à la Savage: define conditions on a preference relation $\succeq$ such that there exists (1) a space of extended states W, (2) a probability distribution P on W and (3) a utility function u such that the preference relation could be rationalized by the expected utility defined over preceding notions.

[15] From [LM81].

(b) Second, the non-standard choice models weakens only the cognitive assumptions of the (underlying) doxastic model. But there remains cognitive assumptions concerning the utility function and the choice criterion. In the approach we just described, we still assume that the agent is able to assign a precise utility to each consequence $c \in C$ and to calculate the solution to its choice criterion. Therefore, from the point of view of the bounded rationality programm, our proposition is strongly incomplete.

4 Conclusion

This paper has advocated the use of non-standard or impossible states as a general framework to "unidealize" belief and choice models. This admittedly does not permit a complete treatment of the idealizations underlying conventional choice models, but can be seen as a first step toward a fine-grained modelling of bounded rationality.

5 Appendix

Proof of Proposition 5

The proof deals only with the case of submodularity ; the other is symmetric. If $[[\phi]]^*$ and $[[\psi]]^*$ are disjoint, then by hypothesis $[[\phi \wedge \psi]]^* = \emptyset$. Therefore $CP(\phi \vee \psi) = CP(\phi) + CP(\psi) - CP(\phi \wedge \psi)$.

It follows from the definition that if $\mathcal{M}, s \vDash \psi \wedge \phi$, then $\mathcal{M}, s \vDash \psi$ and $\mathcal{M}, s \vDash \phi$ (the converse does not hold). In other words,

(1) $[[\phi \wedge \psi]]^* \subseteq [[\phi]]^* \cap [[\psi]]^*$.

This implies that

(2) $P([[\phi \wedge \psi]]^*) \leq P([[\phi]]^* \cap [[\psi]]^*)$.

Since $\mathcal{M}$ is $\vee$-standard, $P([[\phi \vee \psi]]^*) = P([[\phi]]^*) + P([[\phi]]^*) - P([[\phi]]^* \cap [[\psi]]^*)$. By (2), it follows from this that

$P([[\phi \vee \psi]]^*) \leq P([[\phi]]^*) + P([[\phi]]^*) - P([[\phi \wedge \psi]]^*)$. ■

Proof of Theorem 1

($\Rightarrow$). Soundness is easily checked and is left to the reader.

($\Leftarrow$). We have to show that $\models_{NSEPS} \phi$ implies $\vdash_{MPL} \phi$. First, let's notice that the Minimal Probabilistic Logic (MPL) is a "modal logic" ([BdRV01], 191): a set of formulas (1) that contains every propositional tautologies, (2) that is closed by *modus ponens* and uniform substitution. One can then apply the famous Lindenbaum Lemma.

Definition 16

*(i) A formula ϕ is **deducible** from a set of formulas Γ, symbolized $\Gamma \vdash \phi$, if there exists some formulas $\psi_1, ..., \psi_n$ in Γ s.t. $\vdash (\psi_1 \wedge ... \wedge \psi_n) \rightarrow \phi$.*

*(ii) A set of formulas Γ is Λ-**consistent** if it is false that $\Gamma \vdash_\Lambda \bot$*

*(iii) A set of formulas Γ is **maximally Λ-consistent** if (1) it is Λ-consistent and (2) if it is not included in a Λ-consistent set of formulas.*

Lemma 1 (Lindenbaum Lemma)

If Γ is a set of Λ-consistent formulas, then there exists an extension Γ^+ of Γ that is maximally Λ-consistent.

Proof. See for instance [BdRV01], p.199.

Definition 17

Let $\phi \in L$; the language associated with ϕ, $\mathfrak{L}_\phi$ is the smallest sub-language that

(i) contains $\phi, \bot$ et $\top$;

(ii) is closed under sub-formulas

(iii) is closed under the symbol $\sim$ defined as follows : $\sim \chi := \psi$ if $\chi := \neg\psi$ and $\sim \chi := \neg\chi$ if not.[16]

In the language $\mathfrak{L}_\phi$, one can define the analogon of the maximally Λ-consistent sets.

Definition 18
An atom *is a set of formulas in $\mathfrak{L}_\phi$ which is maximally Λ-consistent. $At(\phi)$ is the set of atoms.*

Lemma 2
For every atom Γ,

(i) there exists an unique extension of Γ in L, symbolized Γ^+, that is maximally Λ-consistent ;

(ii) $\Gamma = \Gamma^+ \cap \mathfrak{L}_\phi$

Proof. (i) is an application of Lindenbaum Lemma. (ii) is implied by the fact that Γ is maximally coherent. Suppose that there exists a formula ψ from $\mathfrak{L}_\phi$ in Γ^+ but not in Γ, then Γ^+ would be inconsistent, that is excluded by hypothesis.■

Starting from atoms, one may define the analogon of canonical structures, *i.e.* structures where (standard) states are sets of maximally Λ-consistent formulas. In the same way, we will take as canonical standard state space the language's $\mathfrak{L}_\phi$ atoms.

The hard stuff is the definition of the probability distributions. The aim is to make true in s_Γ every formula $L_a\chi$ in the atom Γ associated with the state s_Γ. To do that, it is necessary that $P(s_\Gamma, \chi) \geq a$; this is guaranted if one takes for $P(s_\Gamma, \chi)$ the number b^* s.t. $b^* = \max\{b : L_b\chi \in \Gamma\}$. This can easily be done with non-standard states. It will be the case if (1) the support of $P(s_\Gamma, .)$ is included in the set of non-standard states, (2) $P(s_\Gamma, .)$ is equiprobable and (3) there is a proportion b^* of states that make χ true.

Suppose that $I(\Gamma)$ is the sequence of formulas in Γ that are prefixed by a doxastic operator L_a ; for every formula, one can rewrite $b^*(\chi)$ as p_i/q_i. Define $q(\Gamma) = \prod_{i\in I} q_i$; $q(\Gamma)$ will be the set of non-standard states in which $P(s_\Gamma, .)$ will be included. If the i-st formula is χ, suffice it to stipulate that χ in the first $p_i \times \prod q_{-i}$ states. One may check that the proportion of states χ is true is p_i/q_i.

Definition 19
*The ϕ-**canonical structure** is the structure $\mathcal{M}_\phi(S_\phi, S'_\phi, \pi_\phi, \vDash_\phi, P_\phi)$ where*

(i) $S_\phi = \{s_\Gamma : \Gamma \in At(\phi)\}$

(ii) $S'_\phi = \bigcup_{\Gamma\in At(\phi)} q(\Gamma)$

(iii) for all standard state, $\pi_\phi(p, s_\Gamma) = 1$ iff $p \in s_\Gamma$

(iv) for all non-standard state s, $\mathcal{M}_\phi, s \vDash_\phi \psi$ iff, if $s \in q(\Gamma)$ and ψ is the i-st formula prefixed by a doxastic operator in Γ, then s is in the $p_i \times \prod q_{-i}$ first states of $q(\Gamma)$

(v) $P_\phi(s_\Gamma, .)$ is an equiprobable distribution on $q(\Gamma)$

As expected, the ϕ-canonical structure satisfies the Truth Lemma.

Lemma 3 (Truth Lemma)
For every atom Γ, $\mathcal{M}_\phi, s_\Gamma \vDash \psi$ iff $\psi \in \Gamma$

Proof. The proof proceeds by induction on the length of the formula.

(a) $\psi := p$; follows directly from the definition of π_ϕ.

(b) $\psi = \psi_1 \vee \psi_2$; by definition, $\mathcal{M}_\phi, s \vDash_\phi \psi$ iff $\mathcal{M}_\phi, s \vDash_\phi \psi_1$ or $\mathcal{M}_\phi, s \vDash_\phi \psi_2$. Case (b) will be checked if one shows that $\psi_1 \vee \psi_2 \in \Gamma$ iff $\psi_1 \in \Gamma$ or $\psi_2 \in \Gamma$. Let's consider the extension Γ^+ of Γ ; one knows that $\psi_1 \vee \psi_2 \in \Gamma^+$ iff $\psi_1 \in \Gamma^+$ or $\psi_2 \in \Gamma^+$. But $\Gamma = \Gamma^+ \cap \mathfrak{L}_\phi$ and ψ_1 and $\psi_2 \in \mathfrak{L}_\phi$. It follows that $s_\Gamma \models \psi_1 \vee \psi_2$ iff $\psi_1 \vee \psi_2 \in \Gamma$.

[16] See [BdRV01], p.242.

(c) $\psi = \neg\chi$. $\mathcal{M}_\phi, s \vDash_\phi \neg\chi$ iff $\mathcal{M}_\phi, s \nvDash_\phi \chi$ iff (by induction hypothesis) $\chi \notin \Gamma$. Suffice it to show that $\chi \notin \Gamma$ iff $\neg\chi \in \Gamma$. (i) Let's suppose that $\chi \notin \Gamma$; χ is in $\mathfrak{L}_\phi$ hence, given the properties of maximally Λ-consistent sets, $\neg\chi \in \Gamma^+$. And since $\Gamma = \Gamma^+ \cap \mathfrak{L}_\phi$, $\neg\chi \in \Gamma$. (ii) Let's suppose that $\neg\chi \in \Gamma$; Γ is coherent, therefore $\chi \notin \Gamma$.

(d) $\psi = L_a\chi$; by definition $s_\Gamma \models L_a\chi$ iff $P(s_\Gamma, \chi) \geq a$. (i) Let's suppose that $P(s_\Gamma, \chi) \geq a$; then $a \leq b^*$ where $b^* = max\{b : L_b\chi \in \Gamma\}$ since by definition of the canonical distribution, $P(s_\Gamma, \chi) = b^*$. Now, let's consider the extension Γ^+ : clearly, $L^*_b\chi \in \Gamma^+$. In virtue of axiom (A7) and of the closure under *modus ponens* of maximally Λ-consistent sets, $L_a \in \Gamma^+$. Given that by hypothesis $L_a\chi \in \mathfrak{L}_\phi$, this implies that $L_a\chi \in \Gamma$. (ii) Let's suppose that $L_a\chi \in \Gamma$; then $a \leq b^*$ hence $P(s_\Gamma, \chi) \geq a$.■

To prove completeness, we need a last lemma.

Lemma 4
Let $At(\phi)$ the set of atoms in $\mathfrak{L}\phi$;
$At(\phi) = \{\Delta \cap \mathfrak{L}\phi : \Delta$ *is maximally coherent* $\}$.

Proof. $At(\phi) \subseteq \{\Delta \cap \mathfrak{L}\phi : \Delta$ is maximally coherent $\}$ follows from a preceding lemma. Let Γ^+ a maximally consistent set and $\Gamma = \Gamma^+ \cap \mathfrak{L}_\phi$. We need to show that Γ is maximally consistent in $\mathfrak{L}_\phi$. First Γ is consistent ; otherwise, Γ^+ would not be. Then, we need to show that Γ is maximal, *i.e.* that for every formula $\psi \in \mathfrak{L}_\phi$, if $\Gamma \cup \{\psi\}$ is consistent, then $\psi \in \Gamma$. Let ψ such a formula. Let's recall that Γ^+ is maximally consistent. Either $\psi \in \Gamma^+$ and then $\psi \in \Gamma$; or $\neg\psi \in \Gamma^+$ (elementary property of maximally consistent sets) and, if $\psi := \neg\chi$, $\chi \in \Gamma^+$ as well. Hence, by definition of $\mathfrak{L}_\phi$, χ or $\neg\neg\chi \in \Gamma$. But this is not compatible with the initial hypothesis according to which $\Gamma \cup \{\phi\}$ is consistent.■

We can now finish the proof : let ϕ a LPM-consistent formula. Then, there exists a maximally LPM-consistent set Γ^+ which contains ϕ. Let $\Gamma = \Gamma^+ \cap \mathfrak{L}_\phi$. ϕ is in Γ therefore by the Truth Lemma, ϕ is true in state s_Γ of the ϕ-canonical structure. Then ϕ is satisfiable. ■

References

[AH02] R.J. Aumann and A. Heifetz. Incomplete information. In R.J. Aumann and S. Hart, editors, *Handbook of Game Theory*, volume 3, pages 1665–1686. Elsevier/North Holland, 2002.

[Aum99] R.J. Aumann. Interactive knowledge. *Internation Journal of Game Theory*, 28:263–300, 1999.

[Bar97] J. Barwise. Information and impossibilities. *Notre Dame Journal of Formal Logic*, 38(4):488–515, 1997.

[BdRV01] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge UP, Cambridge, 2001.

[Coz06] M. Cozic. Epistemic models, logical monotony and substructural logics. In J. van Benthem, editor, *The Age of Alternative Logics*. Kluwer Academic Publisher, 2006.

[FH88] R. Fagin and J.Y. Halpern. Belief, Awareness, and Limited Reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[FH91] R. Fagin and J.Y. Halpern. Uncertainty, Belief and Probability. *Computational Intelligence*, 7:160–173, 1991.

[FHMV95] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, cambridge, Mass., 1995.

[Hac67] I. Hacking. Slightly more Realistic Personal Probability. *Philosophy of Science*, 34:311–325, 1967.

[Hal03] J.Y. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge, Mass., 2003.

[Hin75] J. Hintikka. Impossible worlds vindicated. *Journal of Philosophical Logic*, 4:475–84, 1975.

[HM01] A. Heifetz and P. Mongin. Probability Logic for Type Spaces. *Games and Economics Behavior*, 35:34–53, 2001.

[Lip99] B. Lipman. Decision Theory without Logical Omniscience : Toward an Axiomatic Framework for Bounded Rationality. *The Review of Economic Studies*, 66(2):339–361, 1999.

[LM81] S.A. Lippman and J.J. McCall. The Economics of Uncertainty : Selected Topics and Probabilistic Methods. In K.J. Arrow and M.D. Intriligator, editors, *Handbook of Mathematical Economics, Vol. I*, pages 210–284. Elsevier, 1981.

[LR85] R.D. Luce and H. Raiffa. *Games and Decisions. Introduction and Critical Survey*. Dover, New-York, 2nd edition, 1985.

[Sha76] G. Shafer. *A Mathematical Theory of Evidence*. Princeton UP, Princeton, 1976.

[Sta91] R. Stalnaker. The problem of logical omniscience, i. *Synthese*, 89, 1991.

[Sta98] R. Stalnaker. *Context and Content*. Oxford Cognitive Science. Oxford UP, Oxford, 1998.

Logical Omniscience and Counterpart Semantics

Paul Égré*

Institut Jean-Nicod, CNRS, Paris

Abstract

This paper proposes a pragmatic approach to the phenomenon of hyperintensionality of belief reports in natural language, building on earlier work by J. Gerbrandy (2000) on counterpart semantics for first-order epistemic logic. Counterpart relations are used by Gerbrandy to model the notion of mode of presentation and to account for the context-dependency of *de re* beliefs. I propose to generalize Gerbrandy's semantics to a second-order epistemic logic, in which not only individuals, but also properties and complex relations can have epistemic counterparts. The aim is to give a uniform treatment of cases of hyperintensionality involving expressions of distinct syntactic categories (coreferential proper names, cointensional predicates, logically equivalent sentences), by giving belief sentences a generalized *de re* logical form.

Key Words: *de re* Belief, Counterpart Semantics, Epistemic Logic, Hyperintensionality, Logical Omniscience, Quantified Modal Logic, Pragmatics

The aim of this essay is to propound a new approach to the problem of hyperintensionality of belief contexts, using the apparatus of quantified modal logic and the machinery of counterpart semantics. Consider the following sentence, in which we assume the conditional is a material conditional (but other examples would do):

(1) Peter believes that if door A is locked, then door B is not locked, but he does not believe that if door B is locked, then door A is not locked

In standard epistemic logic (Hintikka 1969) and in intensional logic (Montague 1970), an ascription of belief like this one, which involves two logically equivalent sentences, is predicted to be inconsistent. In Montague grammar, in particular, the proposition expressed by each embedded sentence is the same, and since beliefs are conceived as relations between individuals and propositions, the beliefs of the agents are predicted to be closed under logical equivalence, making them "logically omniscient" in that sense. In modal epistemic logic, the situation is potentially more problematic, since beliefs are predicted to be closed even under logical consequence. There clearly are, however, contexts in which a sentence like (1) above can be uttered consistently. On the other hand, there is also a sense in which the same content is ascribed (or denied) to Peter in either conjunct of (1). The aim of the present proposal is to reconcile these two intuitions, by offering a pragmatic account of hyperintensionality: the basic idea, which I try to articulate in this paper, is that Peter's belief, even though opaque, can be analyzed as a *de re* belief about one and the same proposition (or about the same objects and relation each time), but under different counterpart relations, playing the role of modes of presentation, and acting as a pragmatic component in the evaluation of sentences.

The prime inspiration for this essay comes from the work of J. Gerbrandy (2000) on counterpart semantics for *de re* beliefs. In this paper, I propose to generalize Gerbrandy's semantics to a second order modal logic, in order to account for cases of hyperintensionality involving expressions of distinct syntactic categories (coreferential proper names, cointensional predicates, logically equivalent sentences). Thus the

*This paper is based on chap. 1 of my PhD. dissertation (Égré 2004). Thanks to M. Aloni, D. Bonnay, E. Maier, O. Roy, F. Récanati, P. Schlenker, B. Spector and T. Williamson for helpful comments and criticisms on an earlier draft. Special thanks are due to D. Bonnay for several stimulating and helpful discussions. I also thank audiences in Paris, Amsterdam and Bordeaux, and an anonymous referee for detailed comments and suggestions.

idea is that cases of hyperintensionality should be analyzed on a par with other classic instances of opacity for belief sentences, and the aim is to get a uniform treatment of all those cases. My proposal, however, rests on the idea that belief sentences can be given a *de re* logical form, even in situations which would standardly be analyzed as *de dicto*. This idea raises problems of its own, which will be discussed along the way, but it also contains some potential benefits (like avoiding the resort to impossible worlds beyond standard belief worlds).

The paper is structured as follows. In the first section, I briefly review the *de dicto* treatment of opacity given in Hintikka's modal framework. In section 2, I give some arguments in favor of a *de re* analysis of opacity, and review the treatment of Gerbrandy in section 3. Section 4 presents the generalization of Gerbrandy's semantics to cases of hyperintensionality involving predicates and full sentences. I state a compositional semantics for an appropriate system of second-order modal logic in section 5. The last section offers an evaluation of the benefits and limitations of this proposal, and a comparison with other semantic theories of hyperintensionality.

1 Opacity in epistemic logic

Ordinary belief sentences are notoriously opaque, in the sense that expressions which seem to convey the same information outside the scope of the belief verb are no longer intersubstitutable *salva veritate* under the scope of a verb like "believe". Let us consider the following three examples:

(2) Peter believes that Cicero is a philosopher, but does not believe that Tully is a philosopher.

(3) Peter believes that John is an eye-doctor, but does not believe that John is an ophthalmologist.

(4) Peter believes that if door A is locked, then door B is not locked, but he does not believe that if door B is locked, then door A is not locked.

There is a clear sense in which, in each of these examples, the embedded sentences *say the same thing*, and yet we do find contexts in which each of (2), (3) and (4) can be uttered consistently (which should not be the case if the relevant expressions were always intersubtitutable *salva veritate*). Thus, the proper names "Cicero" and "Tully" are synonymous in so far as they refer to the same individual. Likewise, the predicates "eye-doctor" and "ophthalmologist" are synonymous in so far as they express the same concept. Finally, under the assumption that the conditional is a material conditional in (4), the sentence "if door A is locked, then door B is not locked" and its contrapositive are synonymous in so far as they are logically equivalent. To clarify the problem raised by these sentences, let us represent them in a first-order modal language, letting "$\Box\phi$" stand for "Peter believes that ϕ":

(5) $\Box P(c) \wedge \neg\Box P(t)$

(6) $\Box E(j) \wedge \neg\Box O(j)$

(7) $\Box(L(a) \rightarrow \neg L(b)) \wedge \neg\Box(L(b) \rightarrow \neg L(a))$

Relative to the standard semantics of modal logic ($\Box\phi$ is true at a world if ϕ is true at all accessible worlds), a sentence like (5) is predicted to be inconsistent if one assumes Kripke's thesis of the rigidity of proper names, namely if c and t are taken to denote the same individual in every possible world. Likewise, (6) is inconsistent if one assumes E and O to have the same intension, that is the same extension in each possible world. Finally, $(L(a) \rightarrow \neg L(b))$ and its contrapositive are logically equivalent, true together or false together at every possible world, making (7) inconsistent as well. These examples all illustrate the problem of hyperintensionality of belief contexts: given the standard semantics for epistemic logic, sentences which express the same possible world proposition, that is sentences with the same intension, are predicted to be substitutable *salva veritate* under the scope of a belief operator.[1]

[1]The concept of "hyperintensionality" was introduced by Cresswell 1973, to characterize contexts in which intensionally equivalent sentences are not substitutable *salva veritate*. As a reviewer pointed out, one may distinguish two levels of hyperintensionality, depending on how fine-grained one takes the notion of intensional equivalence to be. Strictly speaking, an operator $\circ$ is *hyperintensional* if there are two formulas ϕ and ψ, a model M and a world w such that $\models (\phi \leftrightarrow \psi)$ (that is ϕ and ψ are logically equivalent), $M, w \models \circ\psi$ and $M, w \not\models \circ\psi$. And following the reviewer's suggestion, one may call an operator *superintensional* if there are two formulas ϕ and ψ, a model M and a world w such that $M \models (\phi \rightarrow \psi)$ (that is ϕ and ψ are model-equivalent), $M, w \models \circ\psi$ and $M, w \not\models \circ\psi$. Hyperintensionality is stronger than superintensionality, since logical equivalence entails model-equivalence. What example (3) suggests is that belief operators are at least superintensional (since the predicates E ("eye-doctor") and O ("ophthal-

The problem is in fact more dramatic in the case of epistemic logic (unlike in intensional logic), since beliefs are predicted to be closed even under logical consequence, due to the monotonicity of the $\Box$ operator. Thus the following sentence is predicted to be inconsistent:

(8) Peter believes that John will come, but does not believe that John or Mary will come.

(9) $\Box C(j) \wedge \neg\Box(C(j) \vee C(m))$

The hyperintensionality of belief sentences therefore suggests that belief verbs shift the ordinary semantic value of sentences. For Hintikka (1969), for instance, two names like "Cicero" and "Tully" are not necessarily coreferential in a belief context simply because an agent can fail to know or realize that they denote the same individual. The same goes for predicates like "eye-doctor" and "ophthalmologist", and also for logically equivalent sentences. Faced with the phenomenon of opacity, it is therefore natural to assume that words can take the meaning they have for the believer, rather than the meaning endorsed by the speaker. Model-theoretically, thus, a sentence like (5) is satisfiable if c and t are allowed to take distinct values in at least one of Peter's belief worlds. Likewise a sentence like (6) is satisfiable if the predicates E and O, although coreferential in the actual world, do not overlap in at least one of Peter's belief worlds. Sentences like (7) and (9) are prima facie more problematic, since the standard epistemic semantics gives logical constants like $\rightarrow$, $\neg$ and $\vee$ a uniform behavior throughout the models.

To go around this problem, a possibility (entertained successively by Montague, Cresswell, Rantala, and advocated by Hintikka 1975) is to enrich the space of worlds with logically impossible worlds, where logically equivalent sentences can take arbitrary values in a non-compositional way. In a way, made explicit by Muskens (1991) and anticipated by Thomason (1980), this solution can be seen as treating logical constants as part of the non-logical vocabulary, and to allow some variation in the denotation of the logical connectives at the belief worlds, in the same way in which constants and predicates are allowed to change their denotation at the belief worlds of the agent. In that manner, a compositional semantics can be given for beliefs, which accounts for opacity at all the relevant syntactic levels. Seen in this light, the problem of *logical omniscience* is ultimately a problem of *semantic competence*: by allowing proper names, predicates, and logical connectives to take on arbitrary values at special worlds, one accounts for the fact that agents can be confused about the objective synonymity of certain expressions.

2 *De re* beliefs and opacity

Despite the coherence of Hintikka's treatment of opacity, there remain several reasons to look for a different solution. A first point of criticism, which has been recurrent on the side of the supporters of Kripke's theory of proper names, concerns the leeway that allows proper names, for instance, to take distinct denotations at the belief worlds of the agent. For a strict Kripkean, proper names are rigid, which means that two names that are coreferential at the actual world should have the same denotation at all the worlds, including the epistemic worlds.[2] A strict supporter of Kripke's theory may allow the predicates "eye-doctor" and "ophthalmologist" to take different values at the belief worlds of Peter, if he grants that those two predicates have a descriptive content, about which Peter can make a mistake. But he will probably disagree with the case of proper names, on the ground that proper names have no descriptive content. This piece of criticism is not entirely convincing, however. For it is one thing to say that proper names are rigid and do not behave like hidden definite descriptions in general, and another thing to acknowledge that, as a matter of plain fact, one can fail to realize that two coreferential proper names are indeed coreferential.[3]

A second, more compelling objection relates to the representation of *de re* beliefs. So far we have given each of the sentences (2)-(4) a *de dicto* logical form, giving the belief operator the largest possible scope (over singular terms, in particular). Several authors, however, have defended the idea that a belief

mologist") need not be equivalent at every world of *every* model – they are model-equivalent only in virtue of a meaning postulate). Example (4), on the other hand, suggests that belief operators are hyperintensional in the strict sense.

[2] Such a view is expressed, in particular, by Recanati (2000: 395), who writes : "According to Hintikka (1962: 138-141), failures of substitutivity in belief contexts show that two co-referential singular terms, though they pick out the same individual in the actual world, may refer to different objects in the ascribee's belief world. That option is ruled out in the present framework; for we want the ontology to be that of the ascriber all along: we want the singular terms to refer to the same objects, whether we are talking of the actual world, or about the ascribee's belief world. That is the price to pay for semantic innocence." *Semantic innocence* is the view according to which an attitude verb does not shift the semantic value of expressions occurring in its scope. This position is criticized in Égré (forthcoming), but I argue, as in what follows, that part of the intuition underlying this view can be reinterpreted.

[3] This is argued quite persuasively by Gerbrandy (2000: 151), and by Aloni (2001: 44-45).

report can be opaque, and nevertheless be *de re* at the same time (Kraut 1983, Heim 1992, Recanati 2000). Recanati, for instance, quoting Loar (1972), insists that "*even on the opaque reading of a belief sentence in which a singular term occurs, reference is made to some particular individual*" (2000, italics his).[4] This claim reflects the following intuition: when one makes a belief ascription like (2), one does not simply mean, according to Recanati, that Peter makes a metalinguistic mistake on the meaning of the proper names "Cicero" and "Tully". In Recanati's theory, the use of two distinct, although coreferential proper names, is just a way of pragmatically indicating that Peter represents to himself one and the same individual under two distinct modes of presentation. When I say in different contexts:

(10) Peter believes that Cicero is a philosopher

(11) Peter does not believe that Tully is a philosopher

I am each time talking about Cicero-Tully, namely about one and the same individual. When I utter the conjunction of the two sentences, I'm still making reference to Cicero-Tully, but the names, being contrasted, pragmatically point to distinct modes of presentation.

Taking this intuition seriously, and considering that a belief can be opaque and nevertheless *de re*, the paraphrase of (2) in modal epistemic logic might therefore be:

(12) $\exists x(x = c \wedge \Box P(x)) \wedge \exists y(y = t \wedge \neg\Box P(y))$

The problem here is that, if c and t denote the same individual d in the actual world (the world of evaluation), the standard semantics for first-order modal logic constrains the variables x and y to denote d across the belief worlds of Peter. Given a first-order model $\langle W, R, D, I\rangle$, an assignment g and a world w, $M, w, g \models \exists x\Box\phi$ if and only if there is a d in D_w such that for every w' such that wRw', $M, w', g[d/x] \models \phi$. This, however, is problematic. First, it makes the ascription a plain contradiction, since the statement then is equivalent to $\exists x(x = c \wedge \Box P(x)) \wedge \exists x(x = c \wedge \neg\Box P(x))$. However, it seems that a sentence like (2) can be uttered without contradiction.

Moreover, it makes a *de re* belief ascription incompatible with situations of mistaken identity, which seems too strong. There are cases where a belief is clearly *de re*, and yet does involve a failure to make a correct identification on the part of the ascribee. The paradigm case is Quine's example of Ralph, who believes of Ortcutt, thought of as "the man seen at the beach", that he is not a spy, and who also believes of Ortcutt, thought of as "the man in the brown hat", that he is a spy. In this scenario, it seems one can make the following belief ascription:

(13) Ralph believes of Ortcutt that he is a spy and Ralph also believes of Ortcutt that he is not a spy.

The conjunction of Ralph's *de re* beliefs can be represented in modal logic by the following sentence, in which the proper names take wide scope over the belief operator:

(14) $\exists x(x = o \wedge \Box S(x)) \wedge \exists x(x = o \wedge \Box\neg S(x))$.

This paraphrase, however, is a problem for the standard semantics of first-order modal logic, since it should then follow that Ralph believes of Ortcutt that he is and is not a spy (see Quine 1956, Aloni 2001), which seems too strong in this context:

(15) $\exists x(x = o \wedge \Box(S(x) \wedge \neg S(x)))$

Quine's example, as argued quite convincingly by Gerbrandy and Aloni, provides a good indication that the standard semantics of first-order modal logic ought to be amended in order to account for *de re* beliefs with mistaken identity. My suggestion here is that roughly the same semantic account which Gerbrandy gives to analyze Quine's example can be extended to parse a sentence like (2) as *de re* instead of *de dicto*. I will first review Gerbrandy's account, then I will attempt to show how to extend this analysis to cases of hyperintensionality involving predicates and full sentences.

[4] By singular term, Recanati means a proper name in the passage under discussion.

3 Counterpart semantics

The two sentences "Ralph believes of Ortcutt that he is a spy", and "Ralph believes of Ortcutt that he is not a spy" are intuitively compatible because the truth of each of them depends on a distinct underlying *mode of presentation*. This mode of presentation need not be explicitly expressed in the sentence, but can be salient to both the speaker and hearer in the discourse situation, so that the two sentences will be seen as mutually compatible. In Gerbrandy's analysis, a mode of presentation is analyzed as a *method of cross-identification*, namely as a way of identifying an individual across epistemic alternatives.[5] Thus one and the same actual individual can have distinct counterparts in the epistemic alternatives of an agent, corresponding to several identification methods. In Quine's scenario, for example, there is one method of identification which "connects objects in Ralph's epistemic alternatives just in case they are the man that Ralph saw...at the beach" (Gerbrandy 2000: 155), and another method which connects objects in Ralph's belief worlds just in case they are the man Ralph saw in the brown hat. Both methods connect objects in Ralph's epistemic alternatives to Ortcutt at the actual world.

In Gerbrandy's semantics, epistemic sentences are evaluated relative to such identity relations, which play the role of a pragmatic parameter. Let us define an epistemic structure as a quadruple $\langle W, R, D, I\rangle$, where W is a set of worlds, R is an epistemic accessibility relation between the worlds, D is a function which associates to each world w a domain of individuals D_w, and I is an interpretation function for the non-logical vocabulary. Formally, a method of identification can be defined as a relation C between ordered pairs (w, d) of worlds and individuals such that $d \in D_w$. If $(w, d)C(w', d')$, d' will be called a counterpart to d in w'. In Gerbrandy's semantics, methods of identification are defined as equivalence relations. A further constraint, which is argued for quite persuasively in Aloni (2001), is to require those relations to be functional, namely such that one individual at a world has at most one counterpart at another world by the identification method, and also to be total, in the sense that if an individual has an epistemic counterpart at a belief world, it has a counterpart at every belief world.[6] Intuitively, the reason why an agent can be mistaken about thinking there are two individuals whereas there is only one, is because this one individual is presented to him in two different ways (there are two different modes of presentation of the same individual). In other words, there should not be more counterparts of an actual individual at a belief world than there are modes of presentation of that individual. But moreover, once a counterpart of an actual individual "inhabits" a belief world, it should also persist at all the belief worlds.

Gerbrandy's semantics for first-order modal logic is standard, except for the epistemic operators and the assignment of variables, which both have to be made sensitive to the identification relations. Thus, given a method of identification C and a pair of worlds w, w', two assignment functions g and h for the variables are in the counterpart relation induced by C if and only if h assigns to every x in w' the counterpart of the individual which g assigns to x in w:

- $g \mapsto_C^{w,w'} h$ iff for every variable x, $(w, g(x))C(w', h(x))$.

The specific satisfaction clauses are the following:

- $M, w, g \models_C \exists x\phi$ iff there is a $d \in D_w$ such that $M, w, g[d/x] \models_C \phi$
- $M, w, g \models_C \Box\phi$ iff for every w' such that wRw', for every h such that $g \mapsto_C^{w,w'} h$: $M, w', h \models_C \phi$.

Given these definitions, Gerbrandy can account for the Ortcutt case. Take a model M with three worlds, where w is the actual world, in which o denotes d, namely Ortcutt, and w' and w'' are the two epistemic alternatives of Ralph. Let us call C_b the identification relation for the beach encounter, and C_h the identification relation for the hat encounter. In w' and w'', d_b is Ortcutt as seen as the beach, namely the counterpart of d under C_b, and in w' and w'', d_h is Ortcutt as seen with the hat, namely the counterpart of d under C_h. Supposing d_b is outside, and d_h is within the denotation of S at both w' and w'', one then has:

(16) $\quad M, w \models_{C_b} \exists x(x = o \wedge \Box\neg S(x))$ & $M, w \models_{C_h} \exists x(x = o \wedge \Box S(x))$

[5] The notion of method of cross-identification originates from Hintikka (1969), and is taken up in Kraut (1983). Much of the inspiration of Gerbrandy's semantics stems from Kaplan (1969) on quantifying in. Counterpart relations originally appear in Lewis (1968).

[6] In what follows, I thus write $C(w, d)(w')$ to denote the counterpart of d in w', relative to w, or $C(d)(w')$ when the world in which d lives is clear.

Using Gerbrandy's counterpart semantics, it is therefore possible to account for Ralph's beliefs, without ascribing a logical contradiction to Ralph.[7]

Now, I would like to suggest that this machinery can be used to give a *de re* analysis of the Tully-Cicero example by which we started. Remember sentence (5), here repeated as (17):

(17) Peter believes that Cicero is a philosopher, but he does not believe that Tully is a philosopher.

We can note that here, unlike in the Ortcutt case, the negation takes wide scope over the belief verb. Formally, this is not a problem, since in the model described for the Ortcutt case, it also holds that:

(18) $M, w \models_{C_b} \exists x(x = o \wedge \neg\Box S(x))$

that is Ralph does not believe that Ortcutt is a spy under the beach identification relation. In the Tully-Cicero example, we can imagine in the same way that Peter's modes of presentations are the names "Tully" and "Cicero": Peter is simply not sure whether those two names of English denote the same individual or not. What this means is that there is at least one epistemic alternative where Cicero-thought-of-as-"Cicero" and Cicero-thought-of-as-"Tully" are distinct individuals. In that situation, the names play the role of methods of identification. It is therefore easy to define two identification relations, namely C_c and C_t, such that, in a three-world model analogous to the previous one, it holds that:

(19) $M, w \models_{C_c} \exists x(x = c \wedge \Box P(x))$ & $M, w \models_{C_t} \exists x(x = t \wedge \neg\Box P(x))$

Using counterpart relations, the standard *de dicto* analysis of a sentence like (17) can therefore be cast into a *de re* analysis, in the spirit of Recanati's suggestions concerning the relational character of opaque belief reports involving proper names. It has actually been suggested that proper names might systematically outscope attitude verbs. This was suggested even for definite descriptions, for totally different reasons, in Heim's analysis of the presupposition projection of attitude verbs (Heim 1992).[8] Other authors, like Kraut (1983), have been even more radical, by defending the idea that there are no *de dicto* attitudes. In what follows, I shall use in a systematic way the idea that *de dicto* belief reports can be restated as *de re* belief reports involving specific modes of presentation or acquaintance relations on the part of the believer.[9]

4 Generalization

In the previous section, we have seen that a certain *de dicto* analysis of substitution failures of proper names can be expressed in terms of a *de re* analysis of substitution failures, using the additional machinery of counterpart semantics. In this section my aim is to show that this analysis can be extended to handle cases of substitution failures involving predicates instead of proper names, and even full sentences, as in the examples by which we started, by allowing higher-order quantification over properties.

This generalization presupposes that it does make sense to talk about *de re* belief about higher-order entities. Two independent arguments can be given for that, however. The first concerns the fact that one can devise scenarios analogous to Quine's Ortcutt case of mistaken identity with properties. Imagine, for instance, a situation in which Peter has two friends, Jack and Jill, having exactly the same profession, namely being eye-doctor, but suppose that Peter is under the misconception that Jack's job is scary (by coincidence, whenever he visits Jack, he sees him perform delicate eye-surgery) while he thinks Jill's job is not (by coincidence, whenever he visits her, he sees her just testing people's eyesight). Unbeknownst to Peter, Jack and Jill perform exactly the same tasks, but at different times. Peter thinks moreover that whoever does the same job has Jack does a scary job, and likewise whoever does the same job as Jill does not do a scary job. As in Quine's example, it seems correct to say:

[7] Gerbrandy gives a similar treatment to Kripke's puzzle of belief (Kripke 1979), in which Pierre is confused about the meaning of the proper names "Londres" and "London".

[8] See Heim 1992: 210-211: "Another way of summarizing the suggestion I just made is this: there is not really just one *de re* reading (for a given constituent), but there are many - one for each acquaintance relation that the context might supply. And some of those many, namely those where the acquaintance relation happens to include the subject's awareness that the *res* fits the same description used by the speaker, are very similar to the *de dicto* reading: more precisely, they entail it. In a way, I am blurring the distinction between *de re* and *de dicto* readings. But that may not be such a bad thing. More often that not, the two are impossible to tell apart."

[9] While doing so, I do not mean to reject the well-foundedness of the *de re-de dicto* distinction in general. An important issue, which I set aside here, concerns the syntactic restrictions that bear on the outscoping of constituents in a sentence, for instance in the case of indefinites.

(20) Peter believes that being an eye-doctor is scary and Peter believes that being an eye-doctor is not scary.

The example invites to treat "being an eye-doctor" as a property about which Peter has two opposite *de re* beliefs. The two conjuncts of (20) may then be analyzed as $\exists X(X = E \wedge \Box Scary(X))$ and $\exists X(X = E \wedge \Box\neg Scary(X))$ respectively.

A second reason to introduce higher-order quantification is that it is needed in order to account for the so-called *non-specific de re* readings of indefinite descriptions in attitude contexts, as in the sentence "Peter believes that some soccer player has a dog", where "some soccer player" is taken *de re* by the speaker, but does not refer to any particular soccer player relative to the believer (see Bonomi 1995). This would happen in a context in which Peter sees a certain dog outside a restaurant, which he thinks belongs to one of the people he saw inside the restaurant, but such that only I, the speaker, know that these people are soccer players.[10] In that case, both the first-order *de dicto* analysis, $\Box\exists x(S(x) \wedge D(x))$, and the first-order *de re* analysis, $\exists x(S(x) \wedge \Box D(x))$ are false, and the correct analysis seems to be: $\exists X(X = S \wedge \Box\exists x(X(x) \wedge D(x)))$, namely "there are soccer players such that Peter believes that one of them has a dog".

Granting the legitimacy of higher-order *de re* beliefs, we are in a position to give a sentence like "Peter believes that John is an eye-doctor" a generalized *de re* logical form, meaning that Peter believes of John and of the property of being an eye-doctor, that the latter applies to the former, namely:

(21) $\exists x\exists X(x = j \wedge X = E \wedge \Box X(x))$

The same analysis can be given for "Peter believes that John is an ophthalmologist", that is: $\exists x\exists X(x = j \wedge X = O \wedge \Box X(x))$. In many contexts, the two attributions will be used interchangeably: for instance, if someone tells me that "Peter believes that John is an eye-doctor", without mentioning anything else about Peter, I may repeat this information to someone else by saying : "Peter believes that John is an ophthalmologist". In such a case, the words are endorsed by the speaker only. In a situation in which I would utter (3), that is "Peter believes that John is an eye-doctor, but does not believe that John is an ophthalmologist", the two *de re* logical forms remain compatible by assuming that each predicate is associated with a distinct counterpart relation, corresponding to a specific acquaintance relation on the part of the believer. Thus one can have:

(22) $M, w \models_{C_E} \exists x\exists X(x = j \wedge X = E \wedge \Box X(x))$
& $M, w \models_{C_O} \exists x\exists X(x = j \wedge X = O \wedge \neg\Box X(x))$

The case of logically equivalent sentences like (4) can be dealt with in the same way, provided one makes room for complex predicates. In the logic we define in the next section, a sentence like "Peter believes that if door A is locked, then door B is not locked" can be given the following generalized *de re* representation:

(23) $\exists x\exists y\exists X(x = a \wedge y = b \wedge X = \lambda xy.(L(x) \rightarrow \neg L(y)) \wedge \Box X(xy))$

This means that Peter believes of door A, of door B, and of the relation such that if one object is locked, then another is not locked, that this relation applies to those individuals. This generalized *de re* analysis allows to use the machinery of counterpart semantics in order to handle the substitution failure of logically equivalent sentences. All it takes is to suppose that the complex relation $\lambda xy.(L(x) \rightarrow \neg L(y))$ has different counterparts in the belief worlds of Peter, depending on the situation in which he is. Thus there might be two methods of identification (for relations), such that:

(24) $M, w \models_{C_1} \exists x\exists y\exists X(x = a \wedge y = b \wedge X = \lambda xy.(L(x) \rightarrow \neg L(y)) \wedge \Box X(xy))$
& $M, w \models_{C_2} \exists x\exists y\exists X(x = a \wedge y = b \wedge X = \lambda xy.(L(y) \rightarrow \neg L(x)) \wedge \neg\Box X(xy))$

Again, one can imagine that these methods of identification are made salient to the hearer by the use of distinct syntactic expressions in each utterance. This reflects the intuition that Peter misperceives the identity of a logical relation between two objects and properties. At the same time, this allows to preserve the idea that Peter's belief, although confused, can very well be a *de re* belief about the objects a and b (for instance in a situation in which he perceives the two doors A and B).

In the same way, closure under logical consequence can be blocked, despite the upward monotonicity of $\Box$, since if $P \subseteq Q$ holds at the actual world, the epistemic counterpart of P need not be a subset of the

[10] I'm indebted to M. Aloni for pointing out Bonomi's example to me. *Non-specific de re* readings have been discussed independently by J. Fodor (1970) and R. Baüerle (1983). See Heim & von Fintel (2002).

epistemic counterpart of Q. For instance, the sentence "Peter believes John will come, but does not believe John or Mary will come", can be paraphrased:

(25) $\exists x \exists y \exists X \exists Y (x = j \wedge y = m \wedge X = C \wedge Y = \lambda xy.(C(x) \vee C(y)) \wedge \Box X(x) \wedge \neg \Box Y(xy))$

The sentence is consistent if one assumes that the counterpart of the relation denoted by Y in the actual world does not hold of John and Mary (assuming those are correctly identified), even though in each of Peter's belief worlds John will come.

5 A second-order epistemic logic

The present section gives the details of the generalization of Gerbrandy's semantics to a second-order modal language. The language, which I call $L_2(\Box)$, is a second-order logic enriched with a unary modal operator (intended as an epistemic modality), in which it is possible to name complex predicates by the usual mechanisms of lambda-abstraction. The present treatment is inspired in part by the presentation of higher-order logic given in the first chapter of Fitting (2002). I only state the semantics here and do not investigate its possible axiomatizations: since the logic is higher-order, we can't expect to have a full completeness result, but rather a Henkin-style completeness proof, but I shall leave this investigation for further work. Another central issue concerns the treatment of quantifiers, and in particular whether we should work with fixed or variable domains (see Fitting & Mendelsohn 1998). The use of counterpart relations makes it natural to let the domains vary, if we think of quantifiers as ranging over objects actually existing at a world, and of counterpart relations as establishing links between distinct ontologies (for instance that of the speaker, and that of the agent whose belief is reported). In what follows we thus place no restriction on the domains (except indirectly, by means of the conditions on counterpart relations), and likewise we give an actualist interpretation to the non-logical vocabulary.

The language $L_2(\Box)$

The definition of $L_2(\Box)$ is in two steps: first I introduce the language L_1 of first-order logic with predicates of arity 0 (propositional symbols), then the notion of a lambda-abstract, which is needed for the definition of the second-order part. The construction is in two steps in order to exclude lambda-abstracts of the form $\lambda x.\Box P(x)$, essentially for reasons of simplicity, so that only non-modal properties can be named. The language is built on an alphabet which includes:

(i) A denumerable set Var of individual variables: $x, y, z, ...$

(ii) For all $n \geq 0$, a denumerable set VAR_n of predicate variables of arity n: $X^n, Y^n, Z^n, ...$ By definition, $VAR = \bigcup_n VAR_n$

(iii) A denumerable set $Cons$ of individual constants: $c, c', c'', ...$

(iv) For all $n \geq 0$, a denumerable set $CONS_n$ of predicate constants of arity n: $P^n, Q^n, R^n, ...$ By definition, $CONS = \bigcup_n CONS_n$.

(v) Logical connectives: $\neg, \wedge, \exists$

(vi) Additional symbols: $\lambda, ., (,)$

(vii) Modality: $\Box$

(viii) Equality symbol: $=$

Individual terms: every element of $Cons$ or Var is an individual term.

L_1-formulas

If t and t' are individual terms, $t = t'$ is a L_1-formula

A predicate constant of arity 0 is a L_1-formula

If $t_1, ..., t_n$ are individual terms, and P is a predicate constant from $CONS_n$, then $P(t_1, ..., t_n)$ is an L_1-formula.

If ϕ and ψ are L_1-formulae, then so are $\neg\phi$ and $(\phi \wedge \psi)$.

If ϕ is an L_1-formula, then $\exists x\phi$ is a L_1-formula

Nothing else is an L_1-formula.

Lambda-abstracts: if ϕ is an L_1-formula, and $x_1, ..., x_n$ are distinct variables from Var, $\lambda x_1, ..., x_n.\phi$ is a lambda-abstract of arity n. We denote by ABS_n the set of lambda-abstracts of arity n, and ABS the set of lambda-abstracts.

If ϕ is an L_1-formula, then $\lambda.\phi$ is a lambda-abstract of arity 0.

$L_2(\Box)$**-formulae**

Every L_1-formula is a $L_2(\Box)$-formula, and if $X \in VAR_0$, X is an $L_2(\Box)$-formula.

If T and T' are respectively a variable in VAR_n, a constant in $CONS_n$, or a lambda-abstract in ABS_n, then $T = T'$ is an $L_2(\Box)$-formula.

If T is a variable in VAR_n, a constant in $CONS_n$, or a lambda-abstract of ABS_n, and $t_1, ..., t_n$ are individual terms, then $T(t_1, ..., t_n)$ is an $L_2(\Box)$-formula.

If ϕ is an $L_2(\Box)$-formula, and X is variable of $VAR_n (n \geq 0)$, then $\exists X\phi$ is an $L_2(\Box)$-formula

If ϕ and ψ are $L_2(\Box)$-formulae, then so are $\neg\phi$ and $(\phi \wedge \psi)$.

If ϕ is an $L_2(\Box)$-formula, then $\Box\phi$ is an $L_2(\Box)$-formula.

Nothing else is an $L_2(\Box)$-formula.

Semantics for $L_2(\Box)$

Model: An $L_2(\Box)$-model M is a quadruple $\langle W, R, D, I\rangle$ where W is a non-empty set ; R is a relation on W ; D is a function which to each world w associates a domain of individuals D_w; I is an interpretation function with domain $W \times (Cons \cup CONS)$, such that: $I_w(c) \in D_w$ if c is an individual constant and $I_w(P) \subseteq (D_w)^n$ for P a predicate constant of arity n.

If P is a predicate symbol of arity 0 (a propositional symbol), then one notes: $I_w(P) = 0$ instead of $I_w(P) = \emptyset$ and $I_w(P) = 1$ instead of $I_w(P) = \{\emptyset\}$. More generally, if D is a set, one notes $D^0 = 1$.

Assignment functions: An assignment function g assigns to a variable x an element in $\bigcup_{w\in W} D_w$, and to a variable X in VAR_n an n-ary relation over $\bigcup_{w\in W} D_w$ (if X has arity 0, again one writes $g(X) = 0$ or $g(X) = 1$).

An assignment g' is an $x_1, ..., x_n$-variant of an assignment g if g and g' give the same values to all variables except at most $x_1, ..., x_n$.

Method of identification: A method of identification C is an equivalence relation between couples (w, d) such that $w \in W$ and $d \in D_w$, and between couples (w, R) such that $w \in W$ and R is an n-ary relation over D_w. Thus the counterpart of an n-ary relation is an n-ary relation. One supposes moreover that C is functional, that is, given (w, d) and w', there is at most one $d' \in D_{w'}$ such that $(w, d)C(w', d')$, and similarly in the case of relations. A further natural constraint is to suppose that C is total in the following sense : if d has a counterpart at one world under C, then it has a counterpart at every other world under C. Thus one can designate by $C(w, d)(w')$ and $C(w, R)(w')$ the respective counterparts in w' of individual d and relation R of w.

Definition: $g \mapsto_C^{w,w'} h$ iff for every variable $x \in Var$ and $X \in VAR$, $(w, g(x))C(w', h(x))$ and $(w, g(X))C(w', h(X))$.

Satisfaction of the formulae

In what follows, I note $T^{(n)}$ to mean that T is a predicate constant, a predicate variable, or a lambda-abstract of arity n. Similarly, $P^{(n)}$ means that P is a predicate constant of arity n, and $X^{(n)}$ means that X is a predicate variable of arity n.

If t is an individual term: one notes $\langle t \rangle_{M,w,g,C} = I_w(t)$ if t is a constant, and $\langle t \rangle_{M,w,g,C} = g(t)$ if t is a variable.

Likewise, one writes $\langle T \rangle_{M,w,g,C} = I_w(T)$ if T is a predicate constant, and $\langle T \rangle_{M,w,g,C} = g(T)$ if T predicate variable.

Finally, if $T = \lambda x_1 ... x_n.\phi$, then $\langle T \rangle_{M,w,g,C} = \{(g'(x_1), ..., g'(x_n)); g'$ is an $x_1, ..., x_n$-variant of g and $g'(x_i) \in D_w$ $(1 \leq i \leq n)$, and $M, w, g' \models_C \phi\}$

If $T = \lambda.\phi$, then $\langle T \rangle_{M,w,g,C} = 1$ if $M, w, g \models_C \phi$ and $\langle T \rangle_{M,w,g,C} = 0$ if $M, w, g \not\models_C \phi$.

$M, w, g \models_C t = t'$ iff $\langle t \rangle_{M,w,g,C} = \langle t' \rangle_{M,w,g,C}$

$M, w, g \models_C T = T'$ iff $\langle T \rangle_{M,w,g,C} = \langle T' \rangle_{M,w,g,C}$

$M, w, g \models_C T^{(0)}$ iff $\langle T \rangle_{M,w,g,C} = 1$, for $T \in VAR_0 \cup CONS_0$.

$M, w, g \models_C T^{(n)}(t_1, ..., t_n)$ iff $(\langle t_1 \rangle_{M,w,g,C}, ..., \langle t_n \rangle_{M,w,g,C}) \in \langle T \rangle_{M,w,g,C}$.

$M, w, g \models_C \neg\phi$ iff $M, w, g \not\models_C \phi$

$M, w, g \models_C (\phi \wedge \psi)$ iff $M, w, g \models_C \phi$ and $M, w, g \models_C \psi$

$M, w, g \models_C \exists x \phi$ iff there exists $d \in D_w$ such that $M, w, g[d/x] \models_C \phi$

$M, w, g \models_C \exists X^{(n)} \phi$ if there exists an n-ary relation $R \subseteq (D_w)^n$ such that $M, w, g[R/X] \models_C \phi$

$M, w, g \models_C \Box\phi$ iff for all w' such that wRw', and for all h such that $g \ \mapsto_C^{w,w'} \ h : M, w', h \models_C \phi$

6 Evaluation

6.1 Comparisons

In principle, the present framework affords the same kind of fine-grainedness that is found in other approaches to hyperintensionality, as in Thomason's treatment in terms of primitive propositions (Thomason 1980), and Muskens' related treatment using impossible worlds (Muskens 1991). The reason is that counterpart relations can be determined by any syntactic component in the embedded sentence. For instance, a sentence like "John believes that if door A is locked then door B is not locked" can be given several *de re* logical forms, as we have seen, including one form in which all the embedded material is scoped out by means of a propositional variable (a variable of arity 0):

(26) $\exists X(X = \lambda.((L(a) \rightarrow \neg L(b)) \wedge \Box X)$

This means that of the proposition that if door A is locked then door B is not locked, Peter believes that it holds. Using appropriate counterpart relations, it is possible to say that Peter believes of that proposition, under one counterpart relation that it holds, and under a different counterpart relation that it does not hold. In Thomason's framework, this corresponds to the fact that the sentences "if door A is locked then door B is not locked" and its contrapositive can very well express different primitive propositions within a model, or be true and false at different non-standard worlds in the case of the impossible worlds approach. However, unlike Thomason or Muskens, our approach of hyperintensionality in terms of counterpart relations does not commit us to a domain of primitive propositions, or of impossible worlds. Ontologically, this is a gain, since all we need are the standard belief worlds of the agent, without having to treat logical constants as non-logical constants.

The closest antecedent to the present approach is Cresswell and von Stechow's (1982) own generalization of the notion of *de re* belief, which they use in particular to represent mathematical beliefs. Thus they observe that the same sentence "Poirot believes that 2+2=4" can be given several distinct logical forms,

depending on the material that is taken *de re* or *de dicto* in the sentence. The present analysis, however, is closer to Hintikka's original motivation, since we see substitution failures first and foremost as failures of the believer to grasp the identity of a property, or of a logical link, as represented by the use of epistemic counterparts.

Also, we do not have to suppose that two syntactically distinct sentences necessarily express different propositions: by default, two logically equivalent sentences will be substitutable in belief contexts, and non-substitutability is seen as a context-dependent phenomenon. This is the sense in which this analysis is fundamentally pragmatic: if ϕ and ψ are intensionally equivalent sentences (in the classic sense), then in some contexts, when uttered after "Peter believes", the clauses "that ϕ" and "that ψ" do express the same information, whereas in other contexts they don't. By giving belief sentences a generalized *de re* logical form, one accounts for the intuition that the belief is about certain entities whose value is determined first relative to the speaker. When I say : "Peter believes that John is an eye-doctor", there is a sense in which I say exactly the same thing as in: "Peter believes that John is an ophthalmologist". In other words, the same literal content is ascribed to Peter. As Recanati argues, situations in which it is appropriate to utter "Peter believes that John is an eye-doctor but does not believe that John is an ophthalmologist" are situations where this content is pragmatically enriched. For Recanati, this pragmatic enrichment does not imply a modification of the basic semantic value of the predicates "ophthalmologist" and "eye-doctor", which should remain constant in the model. The nice feature of Gerbrandy's semantics, in this respect, is that this pragmatic enrichment is materialized by the additional parameter of cross-identification relations, and that the value of lexical terms for which substitution fails needs only to be fixed relative to the speaker. One can always assume, moreover, that the default parameter for contexts where substitution is not at stake is the relation of plain identity. When I utter: "Peter believes that Cicero was poor", without saying more about Peter, the hearer should assume that Peter's belief is about Cicero as commonly identified.

6.2 Refinements

Along with the benefits that we claim for this treatment of hyperintensionality, several imperfections may be pointed out. One of them concerns the treatment we made of conjunction in all the examples we presented so far. This limitation is already present in Gerbrandy's treatment of Quine's example, but is inherited in the other examples we discussed. In order to get a consistent paraphrase of a sentence like (13) above, namely "Ralph believes of Ortcutt that he is a spy, and Ralph believes of Ortcutt that he is not a spy", we have used a metalinguistic conjunction in the form:

(27) $\quad M, w \models_{C_b} \exists x(x = o \wedge \Box S(x))$ & $M, w \models_{C_h} \exists x(x = o \wedge \Box \neg S(x))$

There is no way, in that system, to get a consistent reading of the conjunctive sentence $\exists x(x = o \wedge \Box S(x)) \wedge \exists x(x = o \wedge \Box \neg S(x))$, since sentences are evaluated with respect to only one counterpart relation, which we assumed to be functional. A possible way out would be to relax functionality, but this undermines the intuitive one-to-one correspondence with modes of presentation. Furthermore, it will not be adequate to handle non-conjunctive sentences of mistaken identity like "Peter does not believe that Cicero is Tully", if one analyzes the latter *de re* as $\exists x \exists y(x = c \wedge y = t \wedge \neg \Box(x = y))$.

To go around both problems, another possibility is to allow reference to modes of presentation directly at the sentential level, by indexing variables, as done in Aloni (2001, 2005).[11] In Aloni's system, a sentence like (13), that is "Peter believes that Ortcutt is a spy, and he believes that Ortcutt is not a spy", is represented as: $\exists x_n(x_n = o \wedge \Box S(x_n)) \wedge \exists y_m(y_m = o \wedge \Box \neg S(y_m))$. Likewise, "Peter does not believe that Cicero is Tully" can be paraphrased as "$\exists x_n \exists y_m(x_n = c \wedge y_m = t \wedge \neg \Box(x_n = y_m))$. In Aloni's system, the indices are indices of different conceptual covers. In the same way, we could let the indices denote different counterpart relations supposed to be salient in the discourse context, and evaluate sentences with respect to a family of counterpart relations. Given a *family* C of identification methods, we let C_i denote the identification method indexed by i. We write $g \mapsto_C^{w,w'} h$ iff for every index i and every variables x_i and X_i, $(w, g(x_i))C_i(w', h(x_i))$ and $(w, g(X_i))C_i(w', h(X_i))$. Using this mechanism, one can account for the consistency of mistaken beliefs about identity, and still maintain that the belief is about one and the same actual entity, seen under different modes of presentation.

[11] I am indebted to M. Aloni for these suggestions.

7 Conclusion

In this paper I have defended the view that from a linguistic point of view, that is from the perspective of ordinary belief attributions, the problem of logical omniscience is to be treated on a par with the problems of opacity involving non-logical expressions of other syntactic categories, such as synonymous proper names or predicates, arguing that this problem calls for a pragmatic approach. From the point of view of epistemic logic, the present account can be seen as an extension of Hintikka's original account, since the logic we used allows to give a *de re* representation of sentences that would standardly be analyzed as *de dicto*, and to get a consistent interpretation by means of counterpart relations instead of using impossible worlds. The inspiration remains fundamentally the same, however, since expressions which are synonymous for the speaker are allowed to take arbitrary denotations at the believer's worlds, modulo the counterpart relations. While so doing, we have moved to a much more expressive logic, however. Further work needs to be done, in this respect, to investigate the proof-theoretic properties of the logic, and see what kinds of completeness results can be obtained. More fundamentally, it may be that the best reason we have to hold on to a plain *de dicto* analysis of the examples by which we started is that in all those situations, belief reports have a quotational component without which the ascription simply could not be made. In the present framework, however, this syntactic component is not eliminated, it is simply deferred to the level of counterpart relations.

References

[1] Aloni M. (2001), *Quantification under Conceptual Covers*, ILLC Dissertations Series.

[2] Aloni M. (2005), *Individual Concepts in Modal Predicate Logic*, *Journal of Philosophical Logic*, 34, 1, pp. 1-64 .

[3] Bonomi A. (1995), Transparency and Specificity in Intensional Contexts, in P. Leonardi & M. Santambrogio eds., *On Quine, New Essays*, Cambridge UP: 164-185.

[4] Cresswell M.J. (1973), Hyperintensional Logic, *Studia Logica* 34, pp. 25-38.

[5] Cresswell M.J. & von Stechow A. (1982), *De Re* Belief Generalized, *Linguistics & Philosophy* 5: 503-535.

[6] Égré P. (2004), *Attitudes propositionnelles et paradoxes épistémiques*, Thèse de Doctorat (PhD. Dissertation), Université Paris I Panthéon-Sorbonne, IHPST.

[7] Égré P. (forthcoming), Semantic Innocence and Substitutivity, forthcoming in N. Burton-Roberts, R. Carston et M. J. Frapoli (eds.), *Saying, Meaning and Referring, Essays on François Récanati's Philosophy of Language*, Palgrave Studies in Pragmatics, Language and Cognition.

[8] von Fintel K. & Heim I. (2002), *Lecture Notes on Intensional Semantics*, manuscript, MIT.

[9] Fitting M. & Mendelsohn R.L (1998), *First-Order Modal Logic*, Kluwer.

[10] Fitting M. (2002), *Types, Tableaus, and Gödel's God*, Kluwer, Trends in Logic, Studia Logica Library, vol. 13.

[11] Gerbrandy J. (2000), Identity in Epistemic Semantics, in L. Cavendon, P. Blackburn, N. Braisby & A. Shimojima (eds.), *Logic, Language and Computation*, vol. 3, CSLI, pp. 147-159.

[12] Heim I. (1992), Presupposition Projection and the Semantics of Attitude Verbs, *Journal of Semantics*, 9, pp. 183-221.

[13] Hintikka J. (1969), Semantics for Propositional Attitudes, repr. in L. Linsky (ed.), *Reference and Modality*, Oxford Readings in Philosophy, pp. 145-167.

[14] Hintikka J. (1975), Impossible possible worlds vindicated, *Journal of Philosophical Logic*, 4, pp. 475-484.

[15] Kaplan D. (1969), Quantifying In, repr. in L. Linsky (ed.), *Reference and Modality*, Oxford Readings in Philosophy, pp. 112-144.

[16] Kraut R. (1983), There are no *de dicto* Attitudes, *Synthese*, 54, pp. 275-94.

[17] Kripke S. (1972), *Naming and Necessity*, in D. Davidson & G. Harman (eds), *Semantics for Natural Language*, Reidel.

[18] Kripke S. (1979), A Puzzle about Belief, in A. Margalit (ed), *Meaning and Use*, pp. 239-283, D. Reidel Publishing Co., Dordrecht, Boston.

[19] Lewis D. (1968), Counterpart Theory and Quantified Modal Logic, repr. in M. Loux (ed.), *The Possible and the Actual*, Cornell UP, pp. 11-128.

[20] Loar B. (1972), Reference and propositional attitudes, *Philosophical Review* 81, pp. 43-62.

[21] Montague R. (1970), Pragmatics and Intensional Logic, repr. in R.H. Thomason (ed.), *Formal Philosophy: Selected papers of Richard Montague*, Yale UP, pp. 119-147.

[22] Muskens R. (1991), Hyperfine-Grained Meanings in Classical Logic, *Logique & Analyse*, 133-134, pp. 159-176.

[23] Quine W. V. O. (1956), Quantifiers and Propositional Attitudes, repr. in Linsky (ed.), *Reference and Modality*, Oxford Readings in Philosophy.

[24] Récanati F. (2000a), Opacity and the Attitudes, in Alex Orenstein and Petr Kotatko (eds.), *Knowledge, Language and Logic*, pp. 367-406, Kluwer.

[25] Thomason R. H. (1980), A Model Theory for Propositional Attitudes, *Linguistics & Philosophy*, 4: 47-70.

Rule-based and Resource-bounded: A New Look at Epistemic Logic*

Mark Jago

Abstract

Syntactic logics do not suffer from the problems of logical omniscience but are often thought to lack interesting properties relating to epistemic notions. By focusing on the case of rule-based agents, I develop a framework for modelling resource-bounded agents and show that the resulting models have a number of interesting properties.

1 Introduction

Logical omniscience is a well-documented problem for epistemic logics based on a possible worlds semantics (first presented in Hintikka's seminal *Knowledge and Belief* [21]). In this paper, I concentrate on the concept of belief, as believing ϕ is a necessary condition on knowing ϕ. Belief is defined as truth in all epistemically accessible worlds and as a consequence, belief is closed under consequence and agents automatically believe all valid sentences. This is clearly inadmissible as a general analysis of belief.[1] Several authors take the view that, in a number of situations, logical omniscience is unproblematic, "in particular for interpretations of knowledge that are often appropriate for analyzing distributed systems ... and certain AI systems." However, "it is certainly not appropriate to the extent that we want to model resource-bounded agents" [16, p. 41]. I will therefore take as my starting point the requirement that the beliefs of resource-bounded agents be modelled accurately.

To avoid the problem of logical omniscience, a syntactic approach is required: that is, one which takes the truth-conditions of belief ascriptions to be given, at least in part, in terms of sentences.[2] Contrary to the impression one receives from the logical literature, syntactic accounts of belief receive support from the current philosophical literature.[3] An objection is that syntactic epistemic logics merely give us "ways of *representing* knowledge [and belief] rather than *modelling* knowledge [and belief]". If so, the thought runs, "[o]ne gains very little intuition about knowledge [or belief] from studying syntactic structures" [15, p. 320]. The syntactic approach "lacks the elegance and intuitive appeal of the semantic [possible worlds] approach" [14, p. 40]. My aim in this paper is therefore to present an elegant and intuitively appealing syntactic logic of belief which allows us to accurately model resource-bounded reasoners.

The key idea is to model inference as a nondeterministic step-by-step process. Each time an inference rule is applied and a new belief derived, the agent moves into a new belief state. This is a very fine-grained notion of belief change. It allows models to be built in which perfectly rational reasoning is possible, in the sense that the agent's logical abilities need not be depleted in any way, but in which logical omniscience never arises. This framework models agents that, as [22] has it, are neither logically omniscient nor logically ignorant. The lesson to be taken is that, in order to model real AI agents without making unrealistic assumptions about their resource bounds, an epistemic logic must be able to represent an agent's reasoning at the level of individual inferences (the title of the paper is intended to reinforce this point). My strategy in this paper is to investigate step-by-step inference in a simplified setting. The only inferential action that will be modelled here is the act of deriving new beliefs from old using (a generalised version of) *modus ponens*.

*Thanks to Natasha Alechina for incisive comments, to Brian Logan for guidance and to three anonymous referees for the Logics for Resource Bounded Agents workshop for their helpful suggestions.

[1] See [34, 35, 24] for discussions of logical omniscience and related problems.

[2] Many authors seem to dispute this claim, especially [27, 14, 16, 15]. However, none of the approaches presented there genuinely solve the problem. See [24]. The approach based on *awareness* given in [14] unwittingly concedes the point (see [24].

[3] See Perry [30, 31] and Corazza [11, 10] for accounts of belief in terms of an accepted sentence. Further support comes from accepting the *language of thought hypothesis*: see Fodor [17, 18].

Actions such as making assumptions or instantiating axiom schema are not modelled here (but see [24] in which such actions are modelled in the current framework).

I take as a working example a prominent case from the AI literature: the case of *rule-based agents*. These agents consist of a program—a set of condition-action rules—and a rule interpreter. Rule-based agents have been more or less ignored by the literature on epistemic logic[4] but play an important rôle in other areas of AI. There are several rule-based agent architectures available, e.g. SOAR [26] and SIM-AGENT [33] which allow a great degree of abstraction in specifying behaviour. Rule-based programming extensions are also increasingly being offered as add-ons to existing, lower-level, agent toolkits, e.g., JADE [7] and FIPA-OS [32]. Rule-based behaviour is also playing an important rôle in analysing domains such as business. Business rules (statements that define or constrain an aspect of a business [9], e.g. *every visitor of the conference gets a 20 per cent discount on the first product purchase*) are being used by companies to analyse the behaviour and improve the efficiency of their business. As the business rules community puts it, "business rules are the very essence of a business. They define the terms and state the core business policies. They control or influence business behaviour. They state what is possible and desirable in running a business—and what is not" [9].

In general, a rule-based agent's program will contain condition-action rules of the form

$$\mathtt{P}_1, \ldots, \mathtt{P}_n \Rightarrow \mathtt{Q}_1, \ldots, \mathtt{Q}_m$$

$\mathtt{P}_i$ are the conditions, $\mathtt{Q}_i$ the resulting actions, and each $\mathtt{P}_i$, $\mathtt{Q}_i$ may contain unbound variables or possibly even logical connectives.[5] Here, I treat both rules in the agent's program and literals held in its working memory as beliefs (the working memory does not play a significant rôle in the formalism). In the modal systems discussed below, an agent's rules are represented in the states of those models. An equivalent formulation could be given by encoding rules as conditions on the arcs between states. Intuitively, it makes sense to encode *inference* rules as conditions on arcs and beliefs as the sentences supported by states (the question is whether to treat the rules that appear in the agent's program as inference rules). On the alternative formulation, each rule is treated as an inference rule in its own right whereas on the account presented here, rules are formulae and the agent reasons using a generalised form of *modus ponens*:

$$\frac{\lambda_1, \ldots, \lambda_n, (\lambda_1, \ldots, \lambda_n \Rightarrow \lambda)}{\lambda}$$

In this way, agents are modelled as having many beliefs and only the one rule of inference.

I focus on an agent's reasoning process by assuming that the agent has an initial stock of beliefs (which might be observations) that are neither revised nor added to, other than by firing rules and adding their consequents as new beliefs. I make three further simplifying restrictions: (i) to rules which produce a single action; (ii) to propositional rules and (iii) to rules which contain no disjunctions (thus, on agents who have no disjunctive beliefs). The first two are inessential;[6] (iii) *is* a restriction on the expressiveness of the logic presented here, but is by no means a limitation of the general framework.[7]

The remainder of the paper proceeds as follows. In section 2, I present syntax and semantics for a logic which models a single rule-based agent and then, in section 3, discuss the properties of such models. In section 4, I consider an agent with a fixed program and, in section 5, present an axiomatization and complexity analysis of the resulting logic. Related and future work is discussed in sections 6 and 7.

2 Modelling Rule-Based Agents

We fix a denumerable set of propositions $\mathcal{P} = \{p_1, p_2 \ldots\}$. A literal is either a proposition or its negation; literals are written $\lambda_1, \lambda_2 \ldots$. Rules are of the form $\lambda_1, \ldots, \lambda_n \Rightarrow \lambda$ and in general rules are denoted $\rho, \rho_1, \rho_2, \ldots$. Since it is often useful to know which belief a rule adds when fired, we use the abbreviation

[4] A notable exception is [25]; see section 6.

[5] For example, in the 'definition' rule $\mathtt{person}(x) \Rightarrow \mathtt{man}(x) \vee \mathtt{woman}(x)$.

[6] Because negation may only appear before a predicate—the agent does not believe the negation of any rule—a program in which rules contain unbound variables can be modelled using a denumerable set of propositions, so long as both the set of predicates and the set of constants is denumerable (in any practical case, both will be finite). Using a propositional logic allows us to use a far more readable notation without limiting the underlying logic—all results given below also hold for the predicate case. A logic that deals with predicate-style rules is considered in [4].

[7] Disjunction is ignored here merely to reduce the complexity of the presentation. See [24] for the extended framework, including disjunctions.

$\mathsf{cn}(\rho)$ for λ, given that $\rho = (\lambda_1, \ldots, \lambda_n \Rightarrow \lambda)$. The agent's *internal language* $\mathcal{L}^{\mathcal{P}}$ over $\mathcal{P}$ contains only rules and literals; no other formulae are considered well-formed. Since $\mathcal{P}$ will be fixed throughout, the superscript may be informally dropped. Arbitrary formulae of $\mathcal{L}$ are denoted $\alpha, \alpha_1, \ldots$.

The modal language $\mathcal{ML}^{\mathcal{P}}$, which is used to reason about the agent's beliefs, is built from formulae of $\mathcal{L}^{\mathcal{P}}$ (again the superscript may informally be dropped). $\mathcal{ML}$ contains the usual propositional connectives $\neg, \wedge, \vee, \rightarrow$, the '$\Diamond$' modality and a belief operator B. Given a literal λ and a rule ρ, '$\mathsf{B}\lambda$' and '$\mathsf{B}\rho$' are primitive wffs of $\mathcal{ML}$, and all primitive wffs are formed in this way. If ϕ_1 and ϕ_2 are both $\mathcal{ML}$ wffs, the complex wffs of $\mathcal{ML}$ are then given by

$$\neg\phi_1 \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \phi_1 \rightarrow \phi_2 \mid \Diamond\phi_1$$

The dual modality '$\Box$' is introduced by definition: $\Box\phi \stackrel{df}{=} \neg\Diamond\neg\phi$. Note that the primitive formulae of $\mathcal{ML}$ are all of the form '$\mathsf{B}\alpha$', where 'α' is a $\mathcal{L}$-formula, hence the problem of substitution within belief contexts does not arise in logics based on $\mathcal{ML}$.

Models are graphs of states, with each arc representing a change in an agent's belief state. Although time is not explicitly represented in these models, each arc is thought of as a transition from an agent's belief state at one time to a (possible) belief state at a future moment in time, arrived at by firing a rule and adding its consequent as a new belief. A model M is a structure $\langle S, T, V\rangle$ where S is a set of states; $T \subseteq S \times S$ is a transition relation on states; and $V : S \longrightarrow 2^{\mathcal{L}}$ is the *labelling function*, assigning a set of sentences of the agent's internal language to each state. Where there is a transition from s to s', s' will be said to be a *successor* of s; s' is *reachable* from s when there is a sequence of states $ss_1s_2\cdots s_ns'$ such that each is the successor of the one before.

Definition 1 (Labelling) *Given a model $M = \langle S, T, V\rangle$, a sentence $\alpha \in \mathcal{L}$ is said to* label *a state $s \in S$ when $\alpha \in V(s)$. Given models $M = \langle S, T, V\rangle$ and $M' = \langle S', T', V'\rangle$ (which need not be distinct), states $s \in S$ and $s' \in S'$ are said to be label identical, written $s \leftrightsquigarrow_{\mathcal{L}} s'$, when $V(s) = V'(s')$.*

The definition of a formula ϕ of $\mathcal{ML}$ being true, or satisfied, by state s in a model M (written $M, s \Vdash \phi$) is as follows:

$M, s \Vdash \mathsf{B}\alpha$ iff $\alpha \in V(s)$

$M, s \Vdash \neg\phi$ iff $M, s \nVdash \phi$

$M, s \Vdash \phi_1 \wedge \phi_2$ iff $M, s \Vdash \phi_1$ and $M, s \Vdash \phi_2$

$M, s \Vdash \phi_1 \vee \phi_2$ iff $M, s \Vdash \phi_1$ or $M, s \Vdash \phi_2$

$M, s \Vdash \phi_1 \rightarrow \phi_2$ iff $M, s \nVdash \phi_1$ or $M, s \Vdash \phi_2$

$M, s \Vdash \Diamond\phi$ iff there exists a state $s' \in S$ such that Tss' and $M, s' \Vdash \phi$

Such models are known as Kripke models. '$M, s \Vdash \phi$' is read as *s supports the truth of ϕ in M*, or *s supports ϕ* for short (if it is clear which model is being talked about). The definitions of global satisfiability and validity are standard, and these notion extend to sets for formulae in the usual way. States $s, s' \in S$ are said to be *modally equivalent* in M, written $s \leftrightsquigarrow s'$, when $\{\phi \mid M, s \Vdash \phi\} = \{\psi \mid M, s' \Vdash \psi\}$.

Because these models are common to modal logics in general, they need to be restricted in certain ways to model rule-based agents. In particular, the rules which an agent believes do not change; rules are neither learnt nor forgot. This is standard practise in rule-based AI systems (*cf* condition S4 below). Secondly, T must relate states s and u when some rule ρ can be fired at s, and u is just like s except the agent has gained one new belief, the consequent of ρ. Here, ρ is said to be an s-matching rule.

Definition 2 (Matching rule) *Let ρ be a rule of the form $\lambda_1, \ldots, \lambda_n \Rightarrow \lambda$. ρ is then said to be s-matching, for some state $s \in S$, iff $\rho \in V(s)$, each $\lambda_1, \ldots, \lambda_n \in V(s)$ but $\lambda \notin V(s)$.*

Whenever ρ is s-matching for some state s, then the agent can move into a new state u in which it has gained a new belief. u is said to *extend* s by that new belief, namely $\mathsf{cn}(\rho)$.

Definition 3 (Extension of a state) *For any rule ρ and states $s, u \in S$, u extends s by $\mathsf{cn}(\rho)$ iff $V(u) = V(s) \cup \{\mathsf{cn}(\rho)\}$.*

If there are no matching rules at a state (and so no rule instances to fire), that state is a *terminating state* and has a transition to itself (or to another identical state, which amounts to much the same in modal logic). This ensures that every state has an outgoing transition; in other words, T is a *serial* relation. As a consequence, the question 'what will the agent be doing after n cycles?' may always be answered, even if the agent ran out of rules to fire in less than n cycles.

Definition 4 (Terminating state) *A state s is said to be a terminating state in a model M iff no rule ρ is s-matching.*

Transitions relate terminating states to identically labelled terminating states and, whenever there is a matching rule ρ at a state s, a transition should only be possible to a state u which extends s by $\mathsf{cn}(\rho)$. We capture such transition systems in the class **S** (for single agent models).

Definition 5 *The class **S** contains precisely those models M which satisfy the following:*

S1 *for all states $s \in S$, if a rule $\lambda_1, \ldots, \lambda_n \Rightarrow \lambda$ is s-matching, then there is a state $s' \in S$ such that Tss' and s' extends s by λ.*

S2 *for any terminating state $s \in S$, there exists a state $s' \in S$ such that $V(s') = V(s)$ and Tss'*

S3 *for all states $s, s' \in S$, Tss' only if either* (i) *there is an s-matching rule $\lambda_1, \ldots, \lambda_n \Rightarrow \lambda$ and s' extends s by λ; or* (ii) *s is a terminating state and $V(s) = V(s')$.*

S4 *for all rules ρ and states $s, u \in S$, $\rho \in V(s)$ iff $\rho \in V(u)$.*

It is clear that this definition ensures that T is a serial relation for any model $M \in \mathbf{S}$. For any state $s \in S$, either there is at least one matching rule or there is not. In the former case, S1 ensures that s is related to some extension of itself by T; otherwise, s is a terminating state and is related to an identically labelled state by T.

There may, of course, be many matching rules at a given state, and for each there must be a state u such that Tsu. Each transition may be thought of as corresponding to the agent's nondeterministic choice to fire one of these rule instances. '$\Diamond\phi$' may then be read as 'after some such choice, ϕ will hold.' We can think of the agent's reasoning as a cycle: (i) match rules against literals; (ii) choose one matching rule; (iii) add the consequent of that rule to the set of beliefs; repeat. By chaining diamonds (or boxes), e.g. '$\Diamond\Diamond\Diamond$' we can express what properties can (and what will) hold after so many such cycles. We can abbreviate sequences of n diamonds (or n boxes) as $\Diamond^n$ and $\Box^n$ respectively. '$\Box^n\phi$', for example, may be read as 'ϕ is guaranteed to hold after n cycles.' Note that the agent's set of beliefs grows monotonically state by state and that the agent never revises its beliefs, even if they are internally inconsistent.

Example

Before investigating the properties that models in the class **S** have, an example may help to illustrate the concepts that have been introduced. Typically, the rules in rule-based programs will contain variables which are matched against the contents of the agent's working memory to produce instances of the rule. In this example, the agent's program contains just two rules:

R1 $\mathtt{PremiumCustomer}(x), \mathtt{Product}(y) \Rightarrow \mathtt{Discount}(x, y, 10\%)$

R2 $\mathtt{Spending}(x, > 1000) \Rightarrow \mathtt{PremiumCustomer}(x)$

However, a first-order language is not needed to model this agent. Instead, we can consider the language that contains all instances of the rules and all ground literals that appear in these instances.[8]

Now suppose that the agent's initial working memory contains the beliefs

$$\mathtt{Product}(\mathtt{iBook}) \qquad \mathtt{Spending}(\mathtt{Jones}, > 1000) \qquad \mathtt{Product}(\mathtt{Sunglasses})$$

When the agent begins executing, R2 can be matched against Jones to produce

$$\mathtt{Spending}(\mathtt{Jones}, > 1000) \Rightarrow \mathtt{PremiumCustomer}(\mathtt{Jones}) \tag{1}$$

[8] When considering a program $\mathcal{R}$ (section 4) or the axiomatization given in section 5, we must also assume that the set of constants used to instantiate the variables in rules is finite.

Since no other instances of either R1 or R2 are possible, there is then a unique next state in which

$$\texttt{PremiumCustomer}(\texttt{Jones})$$

is added to the agent's working memory. At the agent's next cycle, R1 can be matched against Jones and either Sunglasses or iBook to produce the instances

$$\texttt{PremiumCustomer}(\texttt{Jones}), \texttt{Product}(\texttt{Sunglasses}) \Rightarrow \texttt{Discount}(\texttt{Jones}, \texttt{Sunglasses}, 10\%) \quad (2)$$

$$\texttt{PremiumCustomer}(\texttt{Jones}), \texttt{Product}(\texttt{iBook}) \Rightarrow \texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%) \quad (3)$$

Note that (1) is no longer counted as a matching rule instance, since its consequent has already been added to the working memory. The agent can then move into a state in which the working memory contains either $\texttt{Discount}(\texttt{Jones}, \texttt{Sunglasses}, 10\%)$ or else contains $\texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%)$ in addition to its previous contents. If the agent fires (2), adding $\texttt{Discount}(\texttt{Jones}, \texttt{Sunglasses}, 10\%)$ to working memory, (3) remains a matching rule instance and $\texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%)$ is added at the next state. Similarly, if the agents fires (3), adding $\texttt{Product}(\texttt{iBook}) \Rightarrow \texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%)$, then (2) remains matching. There is then a next state adding $\texttt{Discount}(\texttt{Jones}, \texttt{Sunglasses}, 10\%)$ to working memory. Figure 1 shows a branching time model in which new beliefs are added to the working memory (only new beliefs are shown). The agent can derive $Discount(Jones, iBook, 10\%)$ in 2 cycles, whereas it must derive it in 3 cycles. If this model is M and its root s, then $M, s \Vdash \Diamond\Diamond\texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%)$ and $M, s \Vdash \Box\Box\Box\texttt{Discount}(\texttt{Jones}, \texttt{iBook}, 10\%)$.

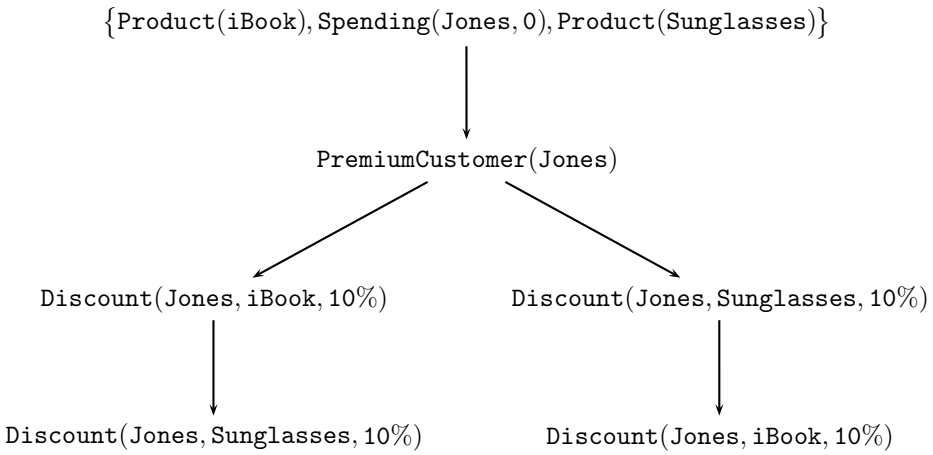


Figure 1: New literals added to WM

3 Properties of Models

Now we need to know, how well do these models capture a rule-based agent's reasoning process? Below I give a number of simple yet powerful results. Firstly, there is a strong relationship between the way states are labelled, the modal formulae which hold at those states and bisimulation. Secondly, models have a *belief convergence* property. The remainder of section is fairly technical.

When a bisimulation relation Z holds between states s, s', we write $s \backsimeq s'$.[9] Intuitively, all bisimilar models describe the same reasoning process. It is sometimes convenient to work with models in which the transition relation T forms a tree on the states S. Such models are known as *tree models*.

Proposition 1 *Some standard properties of models $M = \langle S, T, V\rangle$ and $M' = \langle S', T', V'\rangle$:*

a. For all $s \in S$ and $s' \in S'$, $s \backsimeq s'$ implies $s \leftrightsquigarrow s'$. [8, p.67]

b. Every model M has a bisimilar tree model (obtained by unravelling M).

c. Any satisfiable formula ϕ of depth d is satisfiable in a tree model of height no greater than d.

[9] See, for example, [8] for an explanation of bisimulation.

d. (Hennessy-Milner Theorem) If M and M' are image finite,[10] *then, $s \leftrightsquigarrow s'$ implies $s \backsimeq s'$ for all $s \in S$ and $s' \in S$. [8, p.69]*

These are standard properties of all Kripke models. From (a) and (b), whenever we are working with a model M, we can always switch to a tree model M' which satisfies the same formulae (if $M, s \Vdash \phi$, then there is a state $s' \in M' : M', s' \Vdash \phi$.) The converse to (b) does not hold in general.[11] (c) gives a restricted version of the converse to (b). I now list a few properties which models $M \in \mathbf{S}$ in particular possess (I don't give a proof here as each proof is more or less immediate).

Proposition 2 (Properties of S Models) *Assume that a model $M = \langle S, V, T\rangle$ is a tree model with root r. Then:*

a. For all states s, s' of depth n, $|V(s)| = |V(s')|$. If $V(r)$ is finite and s, s' are not terminating states, then $|V(s)| = n + |V(r)|$.

b. If $V(r)$ is finite, then $V(s)$ is finite for all $s \in S$.

c. If $s \leftrightsquigarrow_{\mathcal{L}} s'$ and s, s' are not terminating states, then s and s' are of the same depth.

d. All siblings of terminating nodes are also terminating nodes.

e. If two children s_1 and s_2 of s are such that $V(s_1) - \{\lambda_1\} = V(s_2) - \{\lambda_2\}$ then each has a child s' such that $V(s') = V(s) \cup \{\lambda_1, \lambda_2\}$.

Lemma 1 *For tree models $M, M' \in \mathbf{S}$ and states s in M, s' in M': if $s \leftrightsquigarrow_{\mathcal{L}} s'$ and Tsu, then there is a $u' \in S'$ such that $Ts'u'$ and $u \leftrightsquigarrow_{\mathcal{L}} u'$.*

Proof: If s is a terminating state then this is trivial; so assume that this is not the case. Then there is an s-matching rule ρ such that $V(u) = V(s)\{\mathsf{cn}(\rho)\}$. Since $s \leftrightsquigarrow_{\mathcal{L}} s'$, ρ is also s'-matching, hence there is a u' such that $Ts'u'$ and $V(u') = V(s') \cup \{\mathsf{cn}(\rho)\}$; hence $u \leftrightsquigarrow_{\mathcal{L}} u'$. ⊣

Theorem 1 *For any models $M, M' \in \mathbf{S}$ and all states s in M and s' in M': $s \leftrightsquigarrow_{\mathcal{L}} s'$ iff $s \leftrightsquigarrow s'$.*

Proof: Clearly, $s \leftrightsquigarrow s'$ implies $s \leftrightsquigarrow_{\mathcal{L}} s'$. The converse: $M, s \Vdash \phi$ iff $M', s' \Vdash \phi$, whenever $s \leftrightsquigarrow_{\mathcal{L}} s'$, is shown by induction on the complexity of ϕ. The base case is trivial so assume that $M, v \Vdash \psi$ iff $M', v' \Vdash \psi$ for all $v \in S$, $v' \in S'$ and ψ of lower complexity than ϕ whenever $v \leftrightsquigarrow_{\mathcal{L}} v'$. The cases for Booleans are also trivial, so consider $\phi := \Diamond\psi$. Then $s \leftrightsquigarrow_{\mathcal{L}} s'$ and $M, s \Vdash \Diamond\psi$ implies that there is a state $u \in S$ such that Tsu and $M, u \Vdash \psi$. By lemma 1, there is a state $u' \in S'$ such that $Ts'u'$ $u' \leftrightsquigarrow_{\mathcal{L}} s'$. By hypothesis, $M', u' \Vdash \psi$ and hence $M', s' \Vdash \Diamond\phi$. The converse holds by a similar argument, hence $s \leftrightsquigarrow s'$. ⊣

Theorem 2 *For any models $M, M' \in \mathbf{S}$ and all states s in M and s' in M': $s \leftrightsquigarrow s'$ iff $s \backsimeq s'$.*

Proof: From proposition 1(a), $s \backsimeq s'$ implies $s \leftrightsquigarrow s'$, so it only remains to show the converse. Assume $s \leftrightsquigarrow s'$ and that there is a $u \in S$ such that Tsu; we must show that there is a state $u' \in S'$ such that $T's'u'$ and $u \leftrightsquigarrow u'$. If s is a terminating state, this is trivial; so assume that s is non-terminating. There must be an s-matching rule ρ. Since $s' \leftrightsquigarrow_{\mathcal{L}} s$, ρ must also be s'-matching and so, by $\mathsf{S1}$, there is a state $u' \in S'$ such that $T's'u'$ extending s' by $\mathsf{cn}(\rho)$. Hence $u' \leftrightsquigarrow_{\mathcal{L}} s'$ and so, by theorem 1, $u' \leftrightsquigarrow s'$. ⊣

Corollary 1 *Let $M = \langle S, T, V\rangle \in \mathbf{S}$. For any $s, s' \in S$ and any descendant u of s, if $s \leftrightsquigarrow_{\mathcal{L}} s'$ then there is a descendant u' of u such that $u \leftrightsquigarrow_{\mathcal{L}} u'$.*

Proof: The proof is immediate from theorem 2. ⊣

We can thus partition states into equivalence classes ($s, s' \in [s]$ whenever $s \leftrightsquigarrow_{\mathcal{L}} s'$) and transform any model M into a bisimular model $M^{\equiv}$ just by comparing the labels on states in M. The domain of $M^{\equiv}$ is the set of label equivalence classes in M such that $T^{\equiv}[s][u]$ whenever Tsu and $V^{\equiv}([s]) = V(s)$ for some $s \in [s]$. Any formula satisfiable in M is then satisfiable in $M^{\equiv}$, and $M^{\equiv}$ has the handy property that $[s] \leftrightsquigarrow_{\mathcal{L}} [u]$ implies $[s] = [u]$.

[10] A model is image finite iff $\bigcup_{s \in S}\{u \mid Tsu\}$ is finite.

[11] Given a model M, we can construct a modally equivalent model N containing an infinite branch for which there can be no bisimulation $Z : M \backsimeq N$ (if we suppose there is, we will eventually come to a point on the infinite branch in N for which the corresponding point in M has no successor; hence they cannot be bisimilar states).

Definition 6 *Let $M = \langle S, T, V\rangle \in \mathbf{S}$ and $n \in \mathbb{N}$. Define $T^n su$ to hold iff there are states $s_0 \cdots s_n$ such that $s = s_0, u = s_n$ and, for each $i < n$, $Ts_n s_{n+1}$.*

Now we show that models in $\mathbf{S}$ have the property of *belief convergence*.

Theorem 3 (Belief Convergence) *For any model $M = \langle S, T, V\rangle \in \mathbf{S}$, any state $r \in S$ and any $n \in \mathbb{N}$, if $T^n rs$ and $T^n ru$, then there is a state s' reachable from s and u' reachable from u such that $s' \leftrightsquigarrow u'$.*

Proof: Without loss of generality, consider a tree model $M \in \mathbf{S}$ whose root is r. Let s, u both be reachable from r in a finite number of transitions. Then there are equinumerous sets X, Y such that $V(s) = V(r) \cup X$ and $V(u) = V(r) \cup Y$. Now consider the subbranch from r to s: for each transition Tvv' on the branch, pick a v-matching rule ρ such that v' extends v by $\mathsf{cn}(\rho)$. Enumerate the selected rules ρ for which $\mathsf{cn}(\rho) \notin V(u)$ as $\rho_1, \ldots, \rho_n$ (from r to s). It is easy to see that there must be a state u' reachable from u, on the branch that results from firing first ρ_1 and then $\ldots$ and then ρ_n. Thus $V(u') = V(u) \cup \{\mathsf{cn}(\rho_1), \ldots, \mathsf{cn}(\rho_n)\} = V(u) \cup X' = V(u) \cup X = V(r) \cup Y \cup X$. By similar reasoning, there must be a state s' reachable from s with $V(s') = V(s) \cup Y = V(r) \cup X \cup Y$. Hence, $s' \leftrightsquigarrow u'$. ⊣

4 Finite Models and Programs

Because of our motivating interest in resource boundedness, we will sometimes want to restrict ourselves to models in which each state is labelled by only finitely many $\mathcal{L}$-formulae, for these are the sentences representing the agent's basic explicit beliefs, of which any real agent may have only finitely many at any one time. We capture this intuition in the class of *finite memory models*.

Definition 7 (Finite memory model) *A model $M \in \mathbf{S}$ is a finite memory model iff $V(s)$ is finite for each $s \in S$. $\mathbf{C}^{fm}$ is the set of all finite memory models in some class $\mathbf{C}$.*

An interesting feature of finite memory models in $\mathbf{S}$ is that each is bisimilar to a finite state model in $\mathbf{S}$. This is the *finite model property*:

Theorem 4 (Finite Model Property) *For any finite memory model $M = \langle S, T, V\rangle \in \mathsf{S}^{fm}$, there is a model M' containing only finitely many states and a bisimulation $Z : M \backsimeq M'$.*

Proof: For any state $s \in S$, if $V(s)$ is finite, s may only have finitely many children, each of which are labelled by only finitely many formulae. Let R be the set of rules which label each state (by S4, all states are labelled by precisely the same rules); clearly R is finite. Then any state $s \in S$ can have at most $|\{\mathsf{cn}(\rho) \mid \rho \in R\}|$ matching rules. Thus a finite memory model with infinitely many states must have an infinite branch, on which only a finite initial segment is generated by matching rules, i.e. only the first n states on the branch are non-terminating states, for some $n \leq |\{\mathsf{cn}(\rho) \mid \rho \in R\}|$. By S3ii, $s \leftrightsquigarrow s'$ whenever Tss' and s, s' are terminating states. A model M' can be obtained by selecting the first terminating state s on each branch in M, removing all the descendants of s and adding a transition Tss. M' satisfies S2 and is clearly bisimilar to M. Moreover, since s occurred on a finite initial segment of a branch in M, M' only contains branches of finite length. It follows that M' only contains finitely many states. ⊣

The above has been a general characterisation of rule-based agents which execute fixed but unspecified set of rules. However, we are often interested in restricting our attention to agents reasoning with a specific set of rules. Following the usual terminology, a *program* is simply a finite set of rules. One of the uses of the current approach is testing for properties of particular programs.[12] Given a program $\mathcal{R}$ for the agent, we can define a subclass $\mathbf{S}_\mathcal{R}$ as containing just those models in $\mathbf{S}$ in which the agent believes all the rules in $\mathcal{R}$ and no further rules.

Definition 8 (The class $\mathbf{S}_\mathcal{R}$) *Let $\mathcal{R}$ be a program (i.e. a finite set of rules). A model $M = \langle S, T, V\rangle \in \mathbf{S}_\mathcal{R}$ iff $M \in \mathbf{S}$ and, for all states $s \in S$, $\mathcal{R} \subseteq V(s)$. An $\mathcal{L}$-formula ϕ is said to be $\mathbf{S}_\mathcal{R}$-satisfiable iff it is satisfied at some state s in some model $M \in \mathbf{S}_\mathcal{R}$.*

[12] In [24] I discuss adding additional temporal operators and path quantifiers from *computational tree logic* (CTL), a common input language for model checking technology. This extension allows rule-based programmers to use current model checking technology to verify their programs.

Each class $\mathbf{S}_\mathcal{R}$ is a subclass of $\mathbf{S}$ and each model in $\mathbf{S}$ is in exactly one class $\mathbf{S}_\mathcal{R}$. $\mathbf{S}$ and its subclasses differ with respect to (semantic) entailment and satisfiability. If $\mathcal{R} = \{p \Rightarrow q\}$, then $\mathsf{B}p \wedge \neg\Diamond\mathsf{B}q$ is $\mathbf{S}$-satisfiable but not $\mathbf{S}_\mathcal{R}$-satisfiable; similarly, $\Diamond\mathsf{B}q$ is a $\mathbf{S}_\mathcal{R}$-consequence but not a $\mathbf{S}$-consequence of $\mathsf{B}p$. The remainder of this section surveys some properties of the class $\mathbf{S}_\mathcal{R}$, including a decidability result.

Theorem 5 *Let $\mathcal{R}$ be a program, ϕ be any $\mathcal{ML}$ formula and $n = |\{\,\mathsf{cn}(\rho) \mid \rho \in \mathcal{R}\}|$. If ϕ is $\mathbf{S}_\mathcal{R}$-satisfiable at all, then it is satisfiable in a finite model $M \in \mathbf{S}_\mathcal{R}$ containing at most n^n states.*

Proof: Suppose ϕ is satisfiable at s in a model in $\mathbf{S}_\mathcal{R}$; then it is satisfied by a tree model $M \in \mathbf{S}_\mathcal{R}$ whose root is s (proposition 1). By S3, any state u in M can have at most $|\mathcal{R}|$ children. Now, take any state s in M of depth n. No $\rho \in \mathcal{R}$ can be s-matching, for otherwise, some ancestor of s must have extended its parent by some $\lambda \notin \{\,\mathsf{cn}(\rho) \mid \rho \in \mathcal{R}\}$; but S3 prohibits this. Then any state at depth n or greater must be a terminating state. There is then a model $M' \in \mathbf{S}_\mathcal{R}$ forming a rooted directed acyclic graph, bisimilar to M, in which $s \leftrightsquigarrow u$ implies $s = u$ (e.g. by taking equivalence classes from M, as described above). For any state s in M', $|\{s' \mid Tss'\}| \leq n$ and, for states u, u' at depth n or greater, $T'uu'$ implies $u = u'$. Therefore M' can contain at most n^n states. ⊣

In any state in a model $M \in \mathbf{S}_\mathcal{R}$, only the labels in the sets $\mathcal{R}$ and $\{\lambda_1, \ldots, \lambda_n, \lambda \mid (\lambda_1, \ldots, \lambda_n \Rightarrow \lambda) \in \mathcal{R}\}$ can have any effect on which rules do and do not match at that state. Thus, it is only these formulae that affect the structure that T forms on S. Labels that are not from these sets may be removed without changing which states are accessible from which in the model. We can combine this with standard techniques to get a notion of filtration for $\mathbf{S}_\mathcal{R}$ models.

Definition 9 ($\mathcal{R}$-filtration) *Let Γ be closed under both subformulae and negation; and set*

$$L_\Gamma \;=\; \mathcal{R} \cup \{\alpha \mid \mathsf{B}\alpha \in \Gamma\} \cup \{\lambda_1, \ldots, \lambda_n, \lambda \mid (\lambda_1, \ldots, \lambda_n \Rightarrow \lambda) \in \mathcal{R}\}$$

An $\mathcal{R}$-filtration of $M = \langle S, T, V\rangle$ through Γ is then a model $M_\Gamma = \langle S, T, V_\Gamma\rangle$ where $V_\Gamma(s) = V(s) \cap L_\Gamma$.

Filtration here is rather different than in regular modal logic. Here, we must ensure that rules and the beliefs needed for them to match are not removed from states when we filter, hence the use of L_Γ.

Lemma 2 *Let Γ be as above, $M = \langle S, T, V\rangle \in \mathbf{S}_\mathcal{R}$ and M_Γ be the $\mathcal{R}$-filtration of M through Γ. Then for any $\phi \in \Gamma$ and $s \in S$: $M, s \Vdash \phi$ iff $M_\Gamma, s \Vdash \phi$.*

Proof: By induction on the complexity of ϕ. If ϕ is an $\mathcal{ML}$ primitive this is trivial. So assume that, for all $\psi \in \Gamma$ of complexity $k < n$ and any state $s \in S$: $M, s \Vdash \psi$ iff $M^L, s \Vdash \psi$. We show this holds for all $\phi \in \Gamma$ of complexity n. The *only if* direction is trivial; in the *if* direction, consider these cases:

$\phi := \neg\psi$. Then $M, s \nVdash \psi$ and, by hypothesis, $M_\Gamma, s \nVdash \psi$, hence $M_\Gamma, s \Vdash \phi$.

$\phi := \psi_1 \wedge \psi_2$. Then $M, s \Vdash \psi_1$ and $M, s \Vdash \psi_2$. By hypothesis, $M_\Gamma, s \Vdash \psi_1$ and $M_\Gamma, s \Vdash \psi_2$, hence $M_\Gamma, s \Vdash \phi$.

$\phi := \Diamond\psi$. Then there is a $s' \in S$ such that $M, s' \Vdash \psi$ and Tss'. By hypothesis, $M_\Gamma, s' \Vdash \psi$ and hence $M_\Gamma, s \Vdash \phi$.

The other Boolean cases are similar; it follows that $M_\Gamma, s \Vdash \phi$. ⊣

Lemma 3 *Let Γ be as above, $M \in \mathbf{S}_\mathcal{R}$ and M_Γ be the $\mathcal{R}$-filtration of M through Γ. Then $M_\Gamma \in \mathbf{S}_\mathcal{R}$.*

Proof: It follows from lemma 2 that any rule ρ is s-matching in M iff it is s-matching in M_Γ and that $\rho \in V_\Gamma(s)$ iff $\rho \in V(s)$. Since T is common to both M and M_Γ, S1-4 are satisfied and hence $M_\Gamma \in \mathbf{S}_\mathcal{R}$.⊣

Definition 10 *Let $sub(\phi)$ be the set of subformulae of ϕ, i.e.:*

$sub(\mathsf{B}\alpha) = \{\mathsf{B}\alpha\}$

$sub(\neg\phi) = sub(\Diamond\phi) = sub(\phi)$

$sub(\phi \wedge \psi) = sub(\phi \vee \psi) = sub(\phi \rightarrow \psi) = sub(\phi) \cup sub(\psi)$

and let $Cl(\phi)$ be $sub(\phi)$ closed under negation.

Theorem 6 (Finite Memory Property) *Let $\mathcal{R}$ be a program and ϕ be any $\mathcal{ML}$ formula. If ϕ is $\mathcal{R}$-satisfiable, then it is satisfiable in a finite memory model $M \in \mathbf{S}_{\mathcal{R}}^{fm}$.*

Proof: Assume that M, s satisfies ϕ. Let M_Γ be the $\mathcal{R}$-filtration of M through $\Gamma = Cl(\phi)$. By lemma 2, $M_\Gamma, s \Vdash \phi$ and, by lemma 3, $M_\Gamma \in \mathbf{S}_{\mathcal{R}}$. Since $Cl(\phi)$ and $\mathcal{R}$ are both finite, $V(s)$ is finite for every $s \in S$, hence $M_\Gamma \in \mathbf{S}_{\mathcal{R}}^{fm}$. ⊣

Theorem 7 (Decidability) *Let $\mathcal{R}$ be a program and ϕ be any $\mathcal{ML}$ formula. Then it is decidable whether ϕ is $\mathbf{S}_{\mathcal{R}}$-satisfiable.*

Proof: Suppose ϕ is $\mathcal{R}$-satisfiable; then it is satisfied at the root r of some tree model $M \in \mathbf{S}_{\mathcal{R}}$. Let M_Γ be the $\mathcal{R}$-filtration of M through $\Gamma = Cl(\phi)$. By inspecting the proof of theorem 6, $M_\Gamma, r \Vdash \phi$, $M_\Gamma \in \mathbf{S}_{\mathcal{R}}^{fm}$ and $V_\Gamma(r) = V(s) \cap L_\Gamma$, with L_Γ as definition 9. Let $n = |\{\, \mathsf{cn}(\rho) \mid \rho \in \mathcal{R}\}|$. By inspecting the proof of theorem 5, a model M'_Γ can be obtained that has at most n^n states (e.g. by taking equivalence classes from M_Γ, as described above). Thus if ϕ has an $\mathbf{S}$-model, one can be found by considering each model with no more then n^n states whose root is labelled by a subse t of L_Γ. Since L_Γ is bounded by the size of ϕ and $\mathcal{R}$, we have an upper bound on the search for a model. We therefore have a terminating algorithm that will find an $\mathbf{S}_{\mathcal{R}}$ model for ϕ if one exists. ⊣

5 Axiomatization and Complexity

Given some such program $\mathcal{R}$, it is easy to axiomatize the logic of the class $\mathbf{S}_{\mathcal{R}}$. The abbreviation

$$\mathsf{match}(\lambda_1, \ldots, \lambda_n \Rightarrow \lambda) \stackrel{df}{=} \mathsf{B}\lambda_1 \wedge \cdots \wedge \mathsf{B}\lambda_n \wedge \neg\mathsf{B}\lambda$$

is helpful. The axiom system shown in figure 2 is called $\Lambda_{\mathcal{R}}$. **A6** says that, when a belief is added, it must have been is added by some matching rule instance in $\mathcal{R}$. **A7** says that, if all matching rule instances in the current state are $\rho_1, \ldots, \rho_n$, then each of the successor states should contain the consequent of one of those instances.

Cl all classical propositional tautologies

K $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$

A1 $\mathsf{B}\rho$ where $\rho \in \mathcal{R}$

A2 $\neg\mathsf{B}\rho$ where $\rho \notin \mathcal{R}$

A3 $\mathsf{B}\alpha \rightarrow \Box\mathsf{B}\alpha$

A4 $\mathsf{B}(\lambda_1, \ldots, \lambda_n \Rightarrow \lambda) \wedge \mathsf{B}\lambda_1 \wedge \cdots \wedge \mathsf{B}\lambda_n \Rightarrow \Diamond\mathsf{B}\lambda$

A5 $\Diamond(\mathsf{B}\alpha \wedge \mathsf{B}\beta) \rightarrow \mathsf{B}\alpha \vee \mathsf{B}\beta$

A6 $\Diamond\mathsf{B}\alpha \rightarrow (\mathsf{B}\alpha \vee \bigvee_{\lambda_1,\ldots,\lambda_n \Rightarrow \lambda \in \mathcal{R}, \lambda = \alpha} \mathsf{B}\lambda_1 \wedge \cdots \wedge \mathsf{B}\lambda_n)$

A7 $\mathsf{match}\rho_1 \wedge \cdots \wedge \mathsf{match}\rho_n \wedge \bigwedge_{\rho \neq \rho_{i \leq n}, \rho \in \mathcal{R}} \neg\mathsf{match}\rho \rightarrow \Box\Big(\mathsf{B}\,\mathsf{cn}(\rho_1) \vee \cdots \vee \mathsf{B}\,\mathsf{cn}(\rho_n)\Big)$ $n > 1$

A8 $\Diamond\top$

MP $\dfrac{\phi \;\; \phi \rightarrow \psi}{\psi}$

N $\dfrac{\phi}{\Box\phi}$

Figure 2: Axiom schemes for $\Lambda_{\mathcal{R}}$

A *derivation* in $\Lambda_{\mathcal{R}}$ is defined in a standard way, relative to $\mathcal{R}$: ϕ is derivable from a set of formulae Γ (written $\Gamma \vdash_{\mathcal{R}} \phi$) iff there is a sequence of formulae $\phi_1, \ldots, \phi_n$ where $\phi_n = \phi$ and each ϕ_i is either an instance of an axiom schema, or a member of Γ, or is obtained from the preceding formulae by **MP** or **N**. Suppose an agent's program $\mathcal{R}$ contains the rules $\rho_1, \ldots, \rho_n$. This agent is guaranteed to reach a state in which it believes α in k steps, starting from a state where it believes $\lambda_1, \ldots, \lambda_m$, iff

$$\mathsf{B}\rho_1 \wedge \ldots \wedge \mathsf{B}\rho_n \wedge \mathsf{B}\lambda_1 \wedge \ldots \wedge \mathsf{B}\lambda_m \Rightarrow \Box^k \mathsf{B}\alpha$$

is derivable in $\Lambda_{\mathcal{R}}$ (again, $\Box^k\alpha$ is an abbreviation for $\Box\Box\cdots\Box\alpha$, k times). We now show that $\Lambda_{\mathcal{R}}$ is the logic of the class $\mathbf{S}_{\mathcal{R}}$ (the proofs of lemmas 4 and 5 are standard).

Lemma 4 (Lindenbaum lemma) *Any set of formulae Γ can be expanded to a $\Lambda_{\mathcal{R}}$-maximal consistent set Γ^+.*

A canonical model $M^{\mathcal{R}} = \langle S, T, V\rangle$ is built in the usual way. States in S are $\Lambda_{\mathcal{R}}$-maximal consistent sets; Tsu iff $\{\phi \mid \Box\phi \in s\} \subseteq u$ (or equivalently, iff $\{\Diamond\phi \mid \phi \in u\} \subseteq s$). Finally, $V(s) = \{\alpha \in \mathcal{L} \mid \mathsf{B}\alpha \in s\}$, for each $s \in S$.

Lemma 5 (Existence and Truth lemma) *For any ϕ and any state s in $M^{\mathcal{R}}$: (i) if there is a formula $\Diamond\phi \in s$ then there is a state u in $M^{\mathcal{R}}$ such that Tsu and $\phi \in u$; and (ii) $M^{\mathcal{R}}, s \Vdash \phi$ iff $\phi \in s$.*

Lemma 6 *Let $M^{\mathcal{R}}$ be a canonical model and let $\alpha \in \mathcal{L}$ and $s, u \in S$. Then (i) if Tsu and $\alpha \in V(u)$ but $\alpha \notin V(s)$, then $V(u) = V(s) \cup \{\alpha\}$; and (ii) α in part (i) must be a literal.*

Proof: Part (i) follows from the definition of '$\Vdash$' together with the truth lemma and the fact that states are closed under axioms **A3** and **A5**. The former axiom ensures that s is a subset of u, the latter ensures that α is the only new belief. For part (ii), if we suppose α were some rule we would have $\alpha \in \mathcal{R}$ and so $\alpha \in s$, contrary to hypothesis. $\dashv$

Theorem 8 (Completeness) *$\Lambda_{\mathcal{R}}$ is strongly complete with respect to the class $\mathbf{S}_{\mathcal{R}}$: given a program $\mathcal{R}$, a set of $\mathcal{ML}$-formulae Γ and an $\mathcal{ML}$-formula ϕ, $\Gamma \Vdash_{\mathcal{R}} \phi$ only if $\Gamma \vdash_{\mathcal{R}} \phi$.*

Proof: Expand Γ to a $\Lambda_{\mathcal{R}}$-maximal consistent set Γ^+ from which we build a canonical model $M^{\mathcal{R}}$. From the truth lemma, it follows that $M^{\mathcal{R}}, \Gamma^+ \Vdash \Gamma$. It remains only to show that $M^{\mathcal{R}}$ is in the class $\mathbf{S}_{\mathcal{R}}$, i.e. that $M^{\mathcal{R}}$ satisfies S1–S4. S4 is clearly satisfied; the remaining cases are:

$M^{\mathcal{R}}$ satisfies S1: Assume there is an s-matching rule ρ. Given the truth lemma, it is easy to see that each of its antecedents is a member of s, whereas its consequent is not. **A4** and the existence lemma guarantee an accessible state u which, given lemma 6, is the extension of s by $\mathsf{cn}(\rho)$.

$M^{\mathcal{R}}$ satisfies S2: Suppose s is a terminating state. By axiom **A8**, there is an accessible state s'. By axiom **A6**, $\alpha \in V(s')$ implies $\alpha \in V(s)$ for any literal α (this holds because there are no matching rules at s). It then follows from axioms **A1**–**A3** that $V(s') = V(s)$, hence S2 is satisfied.

$M^{\mathcal{R}}$ satisfies S3: Suppose Tsu for states s, u in $M^{\mathcal{R}}$. By definition, $\{\phi \mid \Box\phi \in s\} \subseteq u$. By axiom **A7**, there must be one literal believed in u but not in s, namely the consequent of either ρ_1 or $\ldots$ or ρ_n. Then by the argument just used, it follows that u is the extension of s by this new belief. $\dashv$

Theorem 9 *Given a particular program $\mathcal{R}$, the problem of deciding whether a formula ϕ is satisfiable in a model $M \in \mathbf{S}_{\mathcal{R}}$ is NP-complete.*

Proof: Clearly the problem is NP-hard. Let $n = |\{\mathsf{cn}(\rho) \mid \rho \in \mathcal{R}\}|$. From theorem 5, any $\mathbf{S}_{\mathcal{R}}$-satisfiable sentence ϕ has a tree model $M \in \mathbf{S}_{\mathcal{R}}$ containing no more than n^n states which, given the proof of theorem 6, is no larger than $|\phi|n^n$. Given any Kripke structure M', state s in M' and a modal formula ψ, it takes time polynomial in the size of M' and ψ to check whether $M', s \Vdash \psi$ [8]. The crucial point here is that $|\mathcal{R}|$, and hence n^n, is constant in $\mathbf{S}_{\mathcal{R}}$. Thus, we can guess a model $M \in \mathbf{S}_{\mathcal{R}}$ of size no greater than $|\phi|n^n$ and check whether ϕ is satisfied at the root of M in time polynomial in $|\phi|$. It follows that the problem of deciding whether ϕ is $\mathbf{S}_{\mathcal{R}}$-satisfiable is in NP. $\dashv$

One of the main practical uses of models in a class $\mathbf{S}_{\mathcal{R}}$ is to check whether $\mathcal{R}$ satisfies certain properties, specified as an input formula ϕ. One may want to check a range of different programs against a different property: for example, suppose a developer requires an agent which can never move into a state in which ϕ holds. On discovering that ϕ is $\mathbf{S}_{\mathcal{R}_1}$-satisfiable, she must reject $\mathcal{R}_1$. If $\mathcal{R}_2$ is the next generation of the program, then ϕ needs to be checked for $\mathbf{S}_{\mathcal{R}_2}$-satisfiability. The evolution from $\mathcal{R}_1$ to $\mathcal{R}_2$ may have added a large number of rules to the program. This example highlights that it is not just the scalability of satisfiability checking given ϕ as an input that should concern us. How the problem scales with the size of the agent's program is also crucial.[13] An interesting problem to consider, therefore, is the one that takes *both* a formula ϕ and a program $\mathcal{R}$ as its input and determines whether ϕ is $\mathbf{S}_{\mathcal{R}}$ satisfiable. I call this the **S**-SAT problem. The complexity of the problem should be investigated in terms of $|\mathcal{R}|$ and $|\phi|$ rather than in terms of $|\phi|$ alone.

Theorem 10 ***S****-SAT is in PSPACE.*

Proof: The proof is similar to the proof that the K-satisfiability problem has a PSPACE-implementation in [8]. An **S**-Hintikka set over a program $\mathcal{R}$ and a set Σ is like a standard Hintikka set but, in addition, contains all instances of axiom schemes **A1**–**A8** over Σ. A *witness set* is then defined as in [8]. The key result is that a **S**-Hintikka set H over $\mathcal{R}$ and Σ is $\mathbf{S}_{\mathcal{R}}$-satisfiable iff there is a witness set generated by H over $\mathcal{R}$ and Σ. A formula ϕ can then be tested for satisfiability by setting $\Sigma = Cl(\phi) \cup \{\mathsf{B}\alpha \mid \alpha \in Cl(\mathcal{R})\}$. A correct algorithm called *witness* can then be given which returns *true* on input $H, \mathcal{R}, \Sigma$ iff H is a $\mathbf{S}_{\mathcal{R}}$-satisfiable **S**-Hintikka set over $\mathcal{R}$ and Σ. The final stage establishes that *witness* has an implementation on a non-deterministic Turing machine that only requires space polynomial in $|\phi|$ and $|\mathcal{R}|$. Since NPSPACE = PSPACE, this establishes that **S**-SAT is in PSPACE. The full proof is given in [24]. ⊣

6 Related Work

Early work in epistemic logic on rule-based system, influenced by work in AI, is found in Konolige's *Deduction Model of Belief* [25]. As here, semantics is given in terms of sets of formulae, with $\mathsf{B}_i\alpha$ true iff agent i has α in its belief set. Each agent i is assigned a set of deduction rules ρ_i, which need not be logically complete (and in fact must not be to avoid closure of belief under classical consequence). A belief set is then obtained by closing an agent's knowledge base under its rules. This is what [13] term a "final tray" model of belief (p. 1), reporting what an agent *would* derive, given unlimited time and memory. Agents with a functionally complete set of deduction rules are therefore modelled as believing all tautologies and all consequences of their beliefs and so logical omniscience is only avoided by considering agents with depleted logical ability.

In [22, 23], Ho Ngoc Duc presents an epistemic logic based on dynamic logic. If r is an inference rule that the agent can use, then $\langle r \rangle$ is the usual dynamic modality 'after executing (i.e. reasoning using) r, it is possible that ...' where the blank will usually be filled with a belief ascription. Ho introduces a future modality $\langle F \rangle$, defined as the iterated set of all choices of actions $r_1, \ldots, r_n$ available to the agent: $F = (r_1 \cup \cdots \cup r_n)^*$. $\langle F \rangle \mathsf{B}\phi$ then says that the agent can come to believe that ϕ and $[F]\mathsf{B}\phi$ says that the agent must believe that ϕ at some point in the future. The notion of the future here is thus an idealised one, considering all the states in a temporally unbounded reasoning process. For example, if p is a propositional (modality-free) tautology, then $\langle F \rangle \mathsf{B}p$ is a theorem. It is not even correct to read $\langle F \rangle \mathsf{B}p$ as 'the agent can believe p at some point in the (idealised) future' (just consider a tautology p so large that *no* agent could come to hold the sentence in its memory). The $\langle F \rangle$ operator thus ignores resource bounds.

This highlights an important point. Avoiding logical omniscience is not an end in itself. Evidently, what is therefore required is a logic which not only avoids logical omniscience, but that captures the *stages* of reasoning are captured, rather than just the idealised endpoint. Step logic [13] attempts to overcome this problem by indexing beliefs by time points or *steps*. Each step corresponds to a cycle in the agent's reasoning. Step logic deduction rules take the form:

$$\frac{t\text{:} \quad \alpha_1 \ \cdots \ \alpha_n}{t+1\text{:} \qquad \alpha}$$

[13] In fact, this can often be the more important factor of the two, for the size of many programs currently in use far exceeds the size of the formulae that it is useful to check for satisfiability.

However, a semantics is not provided for any step logic in [13]. A minimal possible worlds semantics for step logic are found in [29] and [12]. Belief is defined as a relation between a world and a set of sets of worlds, based on Scott-Montague (or neighbourhood/minimal) structures; an axiomatization is found in [29]. However, agents are modelled as believing all propositional tautologies and their beliefs as closed under equivalence. This is a limitation of Scott-Montague semantics, which deals with the *intensions* of believed sentences (equivalent sentences necessarily have identical intensions). Grant, Kraus and Perlis provide a first-order axiomatization and model theory for step logic in [20]. Not all of the models they describe are adequate representations of an agent's beliefs, in that a particular model may contain 'extra' sentences not derivable from the agent's previous beliefs. Accordingly, they introduce the notion of *knowledge supported models*. This suggests that the framework is not ideally suited to modelling belief obtained by rule-based reasoning.

Timed Reasoning Logic (TRL) is introduced in [5, 6]. The focus is on modelling different rule application and conflict resolution strategies in rule-based systems, building on the step logic approach. Semantics are provided in terms of syntactic *local models*. TRL uses labelled formulae rather than the modal metalanguage adopted here. In [36], TRL is used to model assumption-based reasoning in resource-bounded agents. Such ways of reasoning cannot be modelled by step logic, in which implications must be deal with by forming instances of Hilbert axioms. One major difference between TRL and the present approach is that an agent's current state together with its rules determines a unique next state. It is thus not possible to distinguish between the beliefs that an agent can derive from those it must derive in a certain number of steps. This is a limitation of TRL (and step logic) that has been addressed int he present work.

Ågotnes [1] considers a logic of *finite syntactic epistemic states*. As with TRL and the Deduction Model, the semantics is based on sets of sentences. An unusual feature of [1] is that syntactic operators take *sets* of sentences as their arguments. $\triangle_i\{\phi_1, \ldots, \phi_n\}$ says that agent i believes at least that $\phi_1, \ldots, \phi_n$ are true. Similarly, $\bigtriangledown_i\{\phi_1, \ldots, \phi_n\}$ says that agent i believes at the most that $\phi_1, \ldots, \phi_n$ are true. The syntax of what an agent believes at a time thus closely follows the semantics. A semantics is provided by game-theoretic structures, allowing expressive ATL modalities to be incorporated in the logic. Given a set of agents G, the path quantifier $\langle\!\langle G \rangle\!\rangle$ allows sentences to express co-operation between members of G to achieve some result. This approach forms the basis of [3] and [2].

7 Future Work

This paper has presented a basic framework for modelling rule-based agents in a simplified, monotonic setting. One of the principal applications of the logic that has been developed is to verify that a rule-based program satisfies certain criteria. To this end, the addition of computational tree logic (CTL) modalities would constitute an increase in expressivity and allow the resulting language to be used as an input to model checkers. Note that the $\Diamond$ modality discussed here corresponds to the CTL modality $E\mathcal{X}$ ($E\mathcal{X}\phi$ holds iff ϕ holds at the next step of some branch). This is a minor amendment to the syntax; the models themselves remain identical. The aim in this paper was explicitly to restrict attention to a single rule-based agent. As with most modal logics, it is surprisingly easy to add multiple agents to the formalism (add a valuation V_i for each agent i and plausible rules about communication between agents); see [4].

A more challenging development would be to drop the monotonicity requirement. Nonmonotonic reasoning is important in many areas of AI: see [19]. In fact, a good deal of practical reasoning is nonmonotonic. Makinson comments that "almost all of our everyday reasoning is nonmonotonic; purely deductive, monotonic inference takes place only in rather special contexts, notably pure mathematics" [28, p. 19]. Nonmonotonic reasoning in rule-based systems can arise in a number of ways. One is when certain conditions determine which rule should be fired in the next cycle. Situations can arise in which ρ may fire but, if the agent were to know more information, ρ would not be fired. The resulting consequence relation is nonmonotonic.

Another route to nonmonotonicity in rule-based systems is to consider rules of the form

$$\mathtt{P}_1, \ldots, \mathtt{P}_n \Rightarrow {\sim}\mathtt{Q}$$

where ${\sim}\mathtt{Q}$ instructs the agent to remove $\mathtt{Q}$ from its working memory. Firing such a rule does not lead to a new belief; but it can lead to the agent having one less belief. Amending the current framework to allow for such nonmonotonic rule-based inference would increase its applicability in many areas of AI. A starting point is to amend the requirement that one state extends another when there is a transition to the

first from the second. Instead, define an *amend* operation '$\circ$' on $2^{\mathcal{L}} \times \mathcal{L}$ such that $X \circ p = X \cup \{p\}$ and $X \circ \sim p = X - \{p\}$. Then, whenever there is an s-matching rule ρ, there is a state u such that Tsu and u *amends* s by $\mathsf{cn}(\rho)$. In this system, the order in which rules fire matters. Moreover, it is no longer the case that if Γ entails ϕ then $\Gamma \cup \{\psi\}$ entails ϕ. It would be interesting to see which of the properties discussed above hold of this logic; this is left for future work.

8 Summary

This paper presents a framework for modelling resource bounded reasoners that derive new beliefs from old through inference. The approach is designed to handle inference rules of many types. The example of rule-based programs was chosen here as it allows a simple testbed for the framework. The resulting models of rule-based agents have a number of interesting properties: the equivalence between label identity, modal equivalence and bisimulation, and the belief convergence property. When a particular program is specified, a logic with a decidable satisfaction relation is obtained, which can be easily axiomatized. The interesting satisfiability problem in the resulting logics is in PSPACE.

Not all reasoners are rule-based in the restricted sense used here. Many agents revise their beliefs (and indeed their rules); conclusions can be withdrawn as well as asserted; agents reason inductively and abductively as well as deductively; agents make assumptions and see what follows. These forms of reasoning have not been addressed here. Nevertheless, resource-bounded reasoners using any of these forms of reasoning will amend their set of beliefs in a step-by-step way according to their chosen set of rules. By treating these transitions from one belief state to the next as the foundation for a semantics, a fine-grained account of resource-bounded reasoning is possible in which the problems of logical omniscience never arise.

References

[1] Thomas Ågotnes. *A Logic of Finite Syntactic Epistemic States*. PhD thesis, Department of Informatics, University of Bergen, Norway, April 2004.

[2] Thomas Ågotnes and N. Alechina. The dynamics of syntactic knowledge. Technical Report 304, Dept. of Informatics, Univ. of Bergen, Norway, 2005.

[3] Thomas Ågotnes and M. Walicki. Syntactic knowledge: A logic of reasoning, communication and cooperation. In *Proceedings of the Second European Workshop on Multi-Agent Systems (EUMAS 2004)*, 2004.

[4] N. Alechina, M. Jago, and B. Logan. Modal logics for communicating rule-based agents. In *Proceedings of ECAI 06*, 2006.

[5] N. Alechina, B. Logan, and M. Whitsey. A complete and decidable logic for resource-bounded agents. In *Proc. Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004)*. ACM Press, July 2004.

[6] Natasha Alechina, Brian Logan, and Mark Whitsey. Modelling communicating agents in timed reasoning logics. In *proc. JELIA 04*, pages 95–107, Lisbon, September 2004.

[7] F. Bellifemine, A. Poggi, and G. Rimassa. Developing multi-agent systems with a fipa-compliant agent framework. *Software Practice and Experience*, 21(2):103–128, 2001.

[8] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, New York, 2002.

[9] Business rules community website, accessed 13-03-06. `http://www.brcommunity.com/`, 2006.

[10] Eros Corazza. *Reflecting the Mind: Indexicality and Quasi-Indexicality*. Oxford University Press, 2004.

[11] Eros Corazza. Singular propositions, quasi-singular propositions, and reports. In Kepa Korta, editor, *Semantics, Pragmatics, and Rhetoric*. CSLI, Stanford, 2004.

[12] J. Elgot-Drapkin, S. Kraus, M. Miller, M. Nirkhe, and D. Perlis. Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-4072, University of Maryland, Department of Computer Science, 1999.

[13] J. Elgot-Drapkin, M. Miller, and D. Perlis. Memory, reason and time: the Step-Logic approach. In R. Cummins and J. Pollock, editors, *Philosophy and AI: Essays at the Interface*, pages 79–103. MIT Press, Cambridge, Mass., 1991.

[14] R. Fagin and J.Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[15] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge*. MIT press, 1995.

[16] R. Fagin, J.Y. Halpern, and M.Y. Vardi. A nonstandard approach to the logical omniscience problem. In R. Parikh, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, San Fransisco, California, 1990. Morgan Kaufmann.

[17] Jerry Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, Cambridge, Mass., 1987.

[18] Jerry Fodor. *A Theory of Content and Other Essays*. MIT Press, Cambridge, Mass., 1990.

[19] M. Ginsberg. AI and nonmonotonic reasoning. In D.M. Gabbay *et al*, editor, *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*, pages 1–33. Clarendon Press, Oxford, 1994.

[20] John Grant, Sarit Kraus, and Donald Perlis. A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems*, 2000.

[21] J. Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*. Cornell University Press, Ithaca, N.Y., 1962.

[22] D.N. Ho. Logical omniscience vs. logical ignorance. In C.P. Pereira and N. Mamede, editors, *Proceedings of EPIA'95*, volume 990 of *LNAI*, pages 237–248. Springer, 1995.

[23] D.N. Ho. Reasoning about rational, but not logically omniscient, agents. *Journal of Logic and Computation*, 5:633–648, 1997.

[24] Mark Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, 2006. Forthcoming.

[25] K. Konolige. *A Deduction Model of Belief*. Morgan Kaufman, 1986.

[26] J. E. Laird, A. A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[27] H. J. Levesque. A logic of implicit and explicit belief. In *National Conference on Artificial Intelligence*, pages 1998–202, 1984.

[28] David Makinson. *Bridges from Classical to Nonmonotonic Logic*, volume 5 of *Texts in Computing*. King's College Publications, 2005.

[29] M. Nirkhe, S. Kraus, and D. Perlis. Thinking takes time: a modal active-logic for reasoning in time. Technical Report CS-TR-3249, University of Maryland, Department of Computer Science, 1994.

[30] John Perry. Belief and acceptance. *Midwest Studies in Philosophy*, 5:553–54, 1980.

[31] John Perry. *The Problem of the Essential Indexical*. Oxford University Press, Oxford, 1993.

[32] S. Poslad, P. Buckle, and R. G. Hadingham. The fipa-os agent platform: Open source for open standards. In *Proceedings of the Fifth International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-Agents (PAAM2000)*, pages 355–368, Manchester, April 2000.

[33] A. Sloman and B. Logan. Building cognitively rich agents using the sim agent toolkit. *Communications of the ACM*, 42(3):71–77, March 1999.

[34] R. Stalnaker. The problem of logical omniscience I. *Synthese*, 89:pp.425–440, 1991.

[35] M. Whitsey. Logical omniscience: a survey. Technical Report NOTTCS-WP-2003-2, School of Computer Science and IT, University of Nottingham, 2003.

[36] M. Whitsey. Timed reasoning logics: An example. In *Proc. of the Logic and Communication in Multi-Agent Systems workshop (LCMAS 2004)*. Loria, 2004.

Simulative Inference in a Computational Model of Belief

Aaron Kaplan
Xerox Research Centre Europe

Abstract

Upon learning something new, a person generally draws some conclusions from it, but will miss or ignore some logically valid conclusions (people are not logically omniscient), and furthermore may draw conclusions that do not logically follow from the new information (people's beliefs are not derived purely by deduction). Therefore, in order to reason accurately about what someone believes, it is necessary to know something about how he or she thinks.

We propose a family of logics in which belief is modeled as the result of applying an algorithm to a set of input sentences. We define inference rules that can be used to reason about what others believe by simulating the application of their belief algorithm, and we explore conditions under which these inference rules are sound and complete.

1 Introduction

The human mind is a complex system about whose internal workings we understand very little. Nevertheless, one person can predict with a useful degree of accuracy how another will behave in a wide range of circumstances. It seems plausible that we make such inference by analogy to ourselves, using a kind of introspection: "If I were in such a situation, I would ...." Let us call this type of reasoning *simulative inference*.

In this paper we focus on simulative inference about belief, *i.e.* inference of the form "If I were in a's position, then I would believe ψ; therefore, a believes ψ." In order to formalize this, we posit that beliefs are generated by the application of a reasoning algorithm, and that believers all share the same reasoning algorithm. Simulative inference can then be described as follows:

- Agent a believes that $\varphi_1, \ldots, \varphi_n$.
- If I run my belief algorithm (which is the same one a uses) on background information $\varphi_1, \ldots, \varphi_n$ and query ψ, it answers "yes."
- Therefore, a believes that ψ.

Classical models of belief such as Hintikka's provide a simple axiomatization of belief, but in so doing provide a poor approximation of the real thing, the logical omniscience problem being one notable symptom. Describing belief realistically with a purely axiomatic approach is clearly infeasible—the axiomatization would have to be as complex as the human brain itself. Our approach is thus simply to stipulate that there is some algorithm that believers use for reasoning, and then, without choosing a particular algorithm, to explore the logical properties of belief that result from imposing various constraints on the algorithm. In particular, we will determine a class of algorithms for which simulative inference about belief is sound.

This model is particularly relevant to artificial intelligence applications. If an AI system is endowed with a mechanism intended to simulate how humans reason, then intuitively it seems that the system should be able to reason about humans' beliefs by simulating their thought processes with this mechanism. The logic we present in this paper provides a formal framework for describing this sort of inference, and allows us to define conditions in which it is sound.

2 Overview of main definitions and results

We will now outline the main definitions of our logic and the theorems that follow from them. For the most part we will give only informal descriptions; for details and proofs, the reader is referred to [13].

2.1 Syntax

We define the language L as ordinary first-order logic (FOL) with the addition of a modal belief operator B. Where α is a term and φ is a formula, $B(\alpha, \varphi)$ is a formula whose intended meaning is that α believes φ. "Quantifying in" is allowed, meaning that a variable within a belief formula can be bound by a quantifier outside of that formula, *e.g.* $\exists x.B(a, P(x))$.

2.2 Semantics

Our model of belief is built around the concept of a *belief machine*, which is an abstraction of a computational inference mechanism. In the model, each agent has a belief machine that it uses for storing and retrieving information. The agent enters facts it has learned into its belief machine, and can then pose queries to it. Input and queries are expressed as logical sentences, but the model does not constrain the form in which the machine stores and manipulates the information internally. For example, the machine might use diagrammatic representations, or might have knowledge "hard-coded" into its functional behavior in addition to information it has been given in symbolic form. The machine may perform some inference in answering queries, but it must be guaranteed to give an answer in a finite amount of time. An agent believes a sentence φ if its belief machine is in a state such that the query φ is answered affirmatively.

A belief machine is characterized by two functions, *TELL* and *ASK*. *TELL* describes how the state of the machine changes when a new sentence is stored: if S is the current state of the belief machine, and φ is a sentence, then the value of *TELL*(S, φ) is the new state the belief machine will enter after φ is asserted to it. The value of *ASK*(S, φ) is either yes or no, indicating the response of a machine in state S to the query φ.

Given a particular belief machine m whose set of possible states is Γ, we define the class of m-models as the set of structures $\langle D, I, \gamma \rangle$, where

- D is the domain of individuals,
- I is an interpretation function that maps variables and individual, predicate, and function constants to set-theoretic extensions, as in ordinary FOL,
- $\gamma : D \rightarrow \Gamma$ is a function that assigns each individual a belief state.

A model assigns truth values to formulas of ordinary FOL as usual. A sentence $B(\alpha, \varphi)$ is true in m-model $M = \langle D, I, \gamma \rangle$ if α's belief machine answers yes to φ, *i.e.* when

$$ASK(\gamma(I(\alpha)), \varphi) = yes.$$

We omit the details relating to the semantics of quantification. To summarize, $\exists x.B(a, P(x)) \wedge B(b, Q(x))$ is true in M if there is some individual, and two terms τ_1 and τ_2 that both denote that individual in M, such that $B(a, P(\tau_1))$ and $B(b, Q(\tau_2))$ are both true in M.

2.3 Some metalogical notation

The notation $\mathcal{B} \cdot S$ means the belief set of a machine in state S, *i.e.*

$$\mathcal{B} \cdot S = \{\varphi | \textit{ASK}(S, \varphi) = yes\}.$$

A **sentence** φ is **acceptable** in state S if *TELL*ing the machine φ while it is in state S causes it to believe φ, *i.e.* if

$$\varphi \in \mathcal{B} \cdot \textit{TELL}(S, \varphi).$$

A **sentence** φ is **monotonically acceptable** in state S if it is acceptable in S and *TELL*ing the machine φ while it is in state S does not cause it to retract any beliefs, *i.e.* if

$$\mathcal{B} \cdot S \cup \{\varphi\} \subseteq \mathcal{B} \cdot \textit{TELL}(S, \varphi).$$

A **sequence of sentences** $\varphi_1, \ldots, \varphi_n$ is **monotonically acceptable** in state S if each element φ_i of the sequence is monotonically acceptable in the state *TELL*$(S, \varphi_1, \ldots, \varphi_{i-1})$. A sequence $\varphi_1, \ldots, \varphi_n$ is acceptable in state S_0 (we do not define acceptability of sequences in other states) if

$$\textit{ASK}(\textit{TELL}(S_0, \varphi_1, \ldots, \varphi_n), \varphi_i) = yes$$

for all $1 \leq i \leq n$, and if all initial subsequences of $\varphi_1, \ldots, \varphi_n$ are also acceptable (defined recursively). That is, a sequence is acceptable if, as each of its elements is *TELL*ed to a belief machine starting in the initial state, the machine accepts the new input and continues to believe all of the previous inputs (though it might cease to believe sentences that were inferred from the previous inputs).

2.4 Inference rules

Our logic includes most of the standard inference rules for FOL. Restricted variants are used for the rule of substitution of equals ($B(a, P(b))$ and $b = c$ do not entail $B(a, P(c))$), the rule of $\forall$-substitution ($\forall x B(a, P(x))$ does not entail $B(a, P(c))$), and Skolemization (for some belief machines, $\exists x B(a, P(x))$ is satisfiable yet $B(a, P(k))$ is unsatisfiable—in particular, belief machines that have hard-coded beliefs about k). We also add two new rules for reasoning about belief:

Positive simulative inference: Simulative inference about belief, as we characterized it in the introduction, is formalized by the following natural deduction-style inference rule, where the formulas above the line are the premises, the formula below the line is the conclusion, and the rule can be applied only when the condition written below it holds:

$$\frac{B(\alpha, \varphi_1), \ldots, B(\alpha, \varphi_n)}{B(\alpha, \psi)}$$

if *ASK*(*TELL*$(S_0, \varphi_1, \ldots, \varphi_n), \psi) = yes$.

Although the function *TELL* may be sensitive to the order of its arguments, we will see below that the positive simulative inference rule is only sound for belief machines that satisfy certain constraints; and these constraints entail that if a set of sentences $\varphi_1, \ldots, \varphi_n$ can all be believed simultaneously, then then the order in which they are *TELL*ed is not significant.

Negative simulative inference: In order to have a proof system that is complete (in a restricted sense; see below) with respect to the model theory we have defined, we also need another form of simulative inference, one that allows us to detect by simulation that a set of sentences is not simultaneously believable. As in the former rule, the φ_i in this rule must be sentences, not open formulas; α may be any term.

$$\frac{B(\alpha, \varphi_1), \ldots, B(\alpha, \varphi_n)}{\bot}$$

if *ASK*(*TELL*$(S_0, \varphi_1, \ldots, \varphi_n), \varphi_i) = no$ for some $i, 1 \leq i \leq n$.

This rule says that if *TELL*ing a set of sentences to a belief machine in its initial state doesn't cause it to believe all of those sentences, then no agent can believe all of them simultaneously.

2.5 Soundness and completeness

The inference rules we have defined may or may not be sound, depending on the choice of belief machine. We will state a number of soundness and completeness results that hold for particular classes of belief machines. The classes are defined using various combinations of the following five constraints.

2.5.1 Constraint definitions

Closure: The closure constraint says that *TELL*ing the machine something it already believed does not increase its belief set (though the belief *state* may change).

C1 (closure) *For any belief state S and sentence φ, if*

$$\mathit{ASK}(S, \varphi) = yes$$

then

$$\mathcal{B} \cdot \mathit{TELL}(S, \varphi) = \mathcal{B} \cdot S.$$

This rules out, for example, machines that cut off reasoning after a fixed time or number of inference steps. Positive simulative inference, as we have defined it, is unsound for such machines because in effect they make a distinction between "base beliefs," sentences that have been explicitly *TELL*ed to the machine, and "derived beliefs," those to which the machine assents as a result of inference. Since the logic has only a single belief operator, which describes both base beliefs and derived beliefs, such machines lead to contradictions.

Commutativity: The commutativity constraint says that if a sequence of sentences is acceptable, then it is acceptable in any order, and the belief set of the resulting state does not depend on the order.

C2 (commutativity) *For any belief state S and acceptable sequence of sentences $\varphi_1, \ldots, \varphi_n$, and for any permutation ρ of the integers $1 \ldots n$, the sequence $\varphi_{\rho(1)}, \ldots, \varphi_{\rho(n)}$ is also acceptable, and*

$$\mathcal{B} \cdot \mathit{TELL}(S, \varphi_1, \ldots, \varphi_n) = \mathcal{B} \cdot \mathit{TELL}(S, \varphi_{\rho(1)}, \ldots, \varphi_{\rho(n)}).$$

It is clear that some form of commutativity is necessary for our positive simulative inference rule to be sound, since a pair of premises $B(a, \varphi)$ and $B(a, \psi)$ gives no information about which of φ and ψ came to be believed first. However, the constraint we have stated here stops short of requiring commutativity for all sequences of *TELL*s. It permits the belief machine to take order into account when deciding how to handle input sequences that are not acceptable. Typically these would be sequences in which the machine detects a contradiction.

Monotonicity: This constraint requires that if a *TELL* causes the retraction of some previously held beliefs, then some previously *TELL*ed sentence must be among the retracted beliefs.

C3 (monotonicity) *If a sequence $\varphi_1, \ldots, \varphi_n$ is acceptable, then it is monotonically acceptable.*

Suppose a machine, having been *TELL*ed φ, assents to ψ by default unless it can prove χ. The rule of simulative inference is not sound for such a machine: it licenses the conclusion $B(a, \psi)$ from the premise $B(a, \varphi)$, but the conclusion is not entailed by the premise, since there is a belief state in which φ is believed but ψ is not, namely $\mathit{TELL}(S_0, \varphi, \chi)$. The monotonicity constraint rules out such a machine, since the sequence φ, χ is acceptable but not monotonically acceptable (ψ is believed after the first *TELL*, and no longer believed after the second).

The monotonicity constraint does not completely eliminate the possibility of retraction of beliefs. While it rules out defeasible inference, it does permit machines that, when they discover that their input contains a contradiction, choose some part of the input to ignore.

Acceptable basis: The acceptable basis constraint says that for each belief state, a state with the same belief set can be reached from the initial state by *TELL*ing the machine a finite, acceptable sequence of sentences.

C4 (acceptable basis) *For any belief state S, there exists an acceptable sequence of sentences $\varphi_1, \ldots, \varphi_n$ such that*

$$\mathcal{B} \cdot \mathit{TELL}(S_0, \varphi_1, \ldots, \varphi_n) = \mathcal{B} \cdot S.$$

The monotonicity constraint stated that when retraction occurs, a *TELL*ed sentence must be among the retracted beliefs; the acceptable basis constraint further requires that the effect of the retraction must be the same as that of ignoring part of the input. Note that the effect is not necessarily that of ignoring an entire input sentences—for example, the sequence $\varphi \wedge \psi, \neg\psi$ might induce the belief set $\{\varphi\}$.

Monotonicity under substitution: A belief machine may have "special terms" that it treats differently from others. For example, if a machine has routines for performing mathematical calculations, it may have beliefs about terms like "15" that it doesn't have about terms like c, even in the initial state when it has not yet been *TELL*ed any input sentences. The following constraint requires that even if there are some terms that receive special treatment, the machine still has an infinite supply of "ordinary constants," *i.e.* constants for which the machine has no *a priori* beliefs.

C5 (monotonicity under substitution) *There must be infinitely many individual constants κ such that for any ground term τ and sentences $\varphi_1, \ldots, \varphi_n$,*

1. *if*

$$\mathit{ASK}(\mathit{TELL}(S_0, \varphi_1, \ldots, \varphi_n), \psi) = yes$$

and $\varphi_{1\tau/\kappa}, \ldots, \varphi_{n\tau/\kappa}$ is monotonically acceptable in S_0, then

$$\mathit{ASK}(\mathit{TELL}(S_0, \varphi_{1\tau/\kappa}, \ldots, \varphi_{n\tau/\kappa}), \psi_{\tau/\kappa}) = yes$$

2. *if*

$$\mathit{ASK}(\mathit{TELL}(S_0, \varphi_1, \ldots \varphi_n), \varphi_i) = no$$

for some $1 \leq i \leq n$, then

$$\mathit{ASK}(\mathit{TELL}(S_0, \varphi_{1\tau/\kappa}, \ldots, \varphi_{n\tau/\kappa}), \varphi_{j\,\tau/\kappa}) = no$$

for some $1 \leq j \leq n$.

The intuition behind this constraint is that if one set of sentences contains no less information than another, then the belief machine should draw no fewer conclusions from the first set than from the second. An assertion about a term that has already been used in other assertions, or about a functional term containing function and/or individual constants that have been used in other assertions, conveys more information than the same assertion about a previously unseen constant (such as a Skolem constant). For example, if we know $P(c)$ but know nothing about d, then the assertion $Q(c)$ conveys more information than $Q(d)$, because it makes a connection to other knowledge. The first half of the constraint says that if the belief machine draws the conclusion ψ after being *TELL*ed something about a Skolem constant κ, then it must draw the corresponding conclusion $\psi_{\tau/\kappa}$ if it is *TELL*ed the same thing about any term τ, unless $\psi_{\tau/\kappa}$ is inconsistent with its previous beliefs. The second half of the constraint says that if a sequence of assertions about a Skolem constant is not acceptable, then the sequence must still not be acceptable if the Skolem constant is replaced by any other term.

2.5.2 Soundness and completeness properties

We simply list the soundness and completeness theorems here. For proofs, see [13].

- The positive simulative inference rule is sound for any belief machine that satisfies constraints C1–C4 (closure, commutativity, monotonicity, acceptable basis).
- The negative simulative inference rule is sound for any belief machine that satisfies constraints C1, C3, and C4 (closure, commutativity, and acceptable basis).
- The Skolemization rule is sound for any belief machine that satisfies constraint C5 (monotonicity under substitution).
- The entire logic, which includes the above three rules, is sound for any belief machine that satisfies constraints C1–C5.
- The logic is *not* complete under constraints C1–C5. This is related to the fact that it lacks the compactness property: there are infinite theories that are unsatisfiable yet have no unsatisfiable finite subset.
- The logic is complete for a syntactically restricted subset of the language. Loosely stated, the restriction is that universal quantification into a positively embedded belief context, e.g. $\forall x B(a, P(x))$, is not allowed.

2.6 Introspection and nonmonotonic belief machines

A believer has the property of positive introspection if, for each sentence φ that it believes, it also believes that it believes φ; negative introspection means that when the agent doesn't believe φ, it believes that it doesn't believe φ. If we add an indexical constant me to the logic (which can be interpreted using standard semantic machinery, see [13]), then it becomes possible to construct introspective belief machines: when queried about a sentence of the form $B(me, \varphi)$, a belief machine can simply run the query φ, and return the result of that query. Negative introspection can be implemented similarly.

Though negative introspection is useful and easily implemented, belief machines with this property violate the monotonicity constraint (negative introspection is a form of nonmonotonic reasoning). Thus if we wish to allow negative introspective belief machines, we are forced to discard the rule of positive simulative inference.

Fortunately, the rule of negative simulative inference is sound even for nonmonotonic belief machines; and in fact it turns out to be particularly useful for such machines. Note that if a belief machine is negative introspective, then the following inference rule is sound:

$$\frac{\neg B(\alpha, \varphi)}{B(\alpha, \neg B(me, \varphi))}$$

By chaining this rule with the negative simulative inference rule and *reductio ad absurdum*, we can derive the following new rule, which is sound even for negative introspective machines:

$$\frac{B(\alpha, \varphi_1), \ldots, B(\alpha, \varphi_n)}{B(\alpha, \psi)}$$

if *ASK*(*TELL*$(S_0, \varphi_1, \ldots, \varphi_n, \neg B(me, \psi)), \chi) = no$ for some $\chi \in \{\varphi_1, \ldots, \varphi_n, \neg B(me, \psi)\}$

Note that this derived rule's premises and conclusion are of the same form as those of the original positive simulative inference rule, but the simulation condition is different.

3 Related work

3.1 Possible Worlds Theories

The point of departure for much work on reasoning about belief is the possible worlds model of Hintikka [11], with important refinements by Kripke [18]. Several authors have used the ideas of the possible worlds model, but modified in such a way as to eliminate logical omniscience. In the classical model, a possible world is essentially an ordinary first-order model. Levesque's influential model of "explicit belief" [20], in contrast, uses *situations* which are models of a four-valued logic: a situation can support the truth or falsity of an atomic sentence, or both, or neither. Levesque's proposal has been extended by Lakemeyer [19] to a first-order logic, and by Delgrande [2] to allow nested belief.

In Levesque's model, believers can be seen as perfect reasoners in a weaker-than-ordinary logic which is decidable. Since the logic is decidable, there is a belief machine that implements a sound and complete inference procedure for it, and so there is an instantiation of our logic that is equivalent to Levesque's.

Fagin and Halpern [6] suggest a version of possible worlds semantics in which an agent may have many frames of mind. If an agent believes both φ and ψ, but in different frames of mind, then he believes all of φ's consequences and all of ψ's consequences, but not necessarily any consequences of their combination. The entailments concerning belief are therefore even weaker than those in Levesque's model.

Some authors, for apparently aesthetic reasons, prefer "purely semantic" models like Levesque's to ones like ours in which syntactic entities (sentences of a language of belief) are part of the semantics. But to us, it seems unlikely that belief can be fully explained or described without making reference to the mental representations of the believer, and so it seems entirely appropriate for knowledge representations to be part of a model of belief.

3.2 Sentential Theories

We now turn to models which, like ours, define belief in terms of sentences, and use the concept of inference in the definition of belief.

Our logic can be viewed as a refinement of that of Eberle [4]. That logic has an inferability predicate I, where $I(\varphi_1, \ldots, \varphi_n; \psi)$ means that ψ is inferable from $\varphi_1, \ldots, \varphi_n$. The inferability relation can be an arbitrary relation over sentences, so logical omniscience needn't obtain. An axiom of the logic says that if $\varphi_1, \ldots, \varphi_n$ are believed, and ψ is inferable from $\varphi_1, \ldots, \varphi_n$, then ψ is believed. This is clearly analogous to our rule of positive simulative inference. Eberle also shows that if a constraint analogous to our closure constraint is placed on the inferability relation, then his logic is sound and complete. Our model extends that of Eberle by treating the premises of simulative inference as a sequence rather than a set, and by defining the inferability relation in computational terms.

Halpern, Moses, and Vardi [10] describe a model of "algorithmic knowledge" that is similar to our model in that an agent is modeled has having a reasoning algorithm, and its beliefs are taken to be the sentences that the algorithm can verify given the information encoded in the agent's knowledge state. Simulative inference has not been discussed in the framework of algorithmic knowledge. In addition, Halpern *et al*'s logic is a propositional logic, whereas ours is first-order. Soundness and completeness proofs are significantly more complicated in the first-order case.

The mode of reasoning described by Haas [8, 9] is also simulative reasoning, but of a different kind. It can be summarized as follows: if α believes φ at time t, and I can prove ψ from φ in n time units, then α believes ψ by time $t + n$. Since beliefs are indexed with a time at which they are known to be believed, this simulative technique is sound for a broader class of reasoning mechanisms than ours. Specifically, the closure constraint is not required for soundness, because the logic differentiates between base beliefs and derived beliefs. However, implicit in Haas' presentation is the assumption that belief computation is simply the enumeration of conclusions that can be reached from a set of premises, not goal-directed inference that attempts to verify or refute a particular query sentence. If the mechanism being used in simulative reasoning were a goal-directed theorem prover, then one would not be justified in concluding that another agent with similar inferential ability had reached the same conclusion in the same amount of time. For simulation to be justified, one would have to know, in addition, that the agent had had occasion to "wonder" about the query sentence, *i.e.* had begun attempting to prove it.

Related ideas of time-bounded inference can be found in the "step logics" (later renamed "active logics") of Elgot-Drapkin *et al.* [5]. In these logics, an agent's beliefs are seen as a set of sentences that changes over time, at integral time steps, as a result of both observation and inference. Inference can introduce new sentences derived (via sound or unsound rules) from sentences believed in the previous time step, and can also cause the retraction of beliefs held in the previous step, as a result of contradictions discovered in the earlier belief set. It may well be possible to develop a form of simulative inference in this framework, but Elgot-Drapkin *et al.* have not done so. Instead, they deal with reasoning about others' beliefs by using axiomatic descriptions of belief, *e.g.* an axiom that says "if a believes φ and $\varphi \rightarrow \psi$ at time t, then he believes ψ at time $t+1$." Note the difference between this rule and Haas' simulative inference rule described in the previous paragraph: Elgot-Drapkin *et al.*'s reasoning paradigm requires one application of the belief rule for every step of the proof being attributed to the other agent, whereas a simulative inference rule attributes an entire chain of reasoning to the other agent in a single step.

The logics of Haas [9], Perlis [21], and Grove [7] have in common that they allow explicit quantification over names. In our model, in contrast, while quantifying-in is implicitly quantification over names, terms of the logic are not elements of the domain of discourse. As a result, our logic is not quite as expressive; Grove gives some examples of situations in which reasoning explicitly about names is desirable. The tradeoff is that both the syntax and the inference methods for our logic are significantly simpler.

3.3 Temporal and Dynamic Logics

The logic we have presented here is a static one, in the sense that it can only be used for reasoning about what an agent believes at one moment in time. The aforementioned logics of Haas and of Elgot-Drapkin *et al.*, in contrast, are temporal logics in which one can make assertions about what an agent believes at different times, and thus about how its beliefs change in response to new information. Other logics allow reasoning about belief change without an explicit representation of points in time, by introducing epistemic operators with meanings such as, "after a learns that φ, ψ will hold" [22, 24, 23], or "a will believe φ at some future time" [3, 1]. The logics in the first group, the family of "dynamic epistemic logics" represented *e.g.* by van Ditmarsch *et al.* [24], is based on standard Kripke semantics, with its idealized "logically omniscient" believers. The logics in the second group, those of Ho Ngoc Duc [3], including the logic later elaborated by Ågnotes and Alechina [1], avoid this idealization, but since their epistemic

operators do not allow one to express that an agent will believe a proposition at a particular point in time, they are less useful for applications such as reasoning about games. Pucella's logic is an extension of that of Konolige [17], to which our logic bears important similarities; we devote the entire next section to a comparison with Konolige's logic.

Although the logic we have presented in this paper lacks the syntax that these temporal and dynamic logics provide to support reasoning about belief change, the computational model of belief that underlies it is built around notions of belief state and belief change, and so would lend itself quite naturally to use in a dynamic logic. We view this as a promising direction for future work.

3.3.1 Konolige's Deduction Model

In its essence, our model of belief bears a strong resemblance to Konolige's deduction model [17]. In the deduction model, a believer is represented by a *deduction structure*, which is composed of a set of sentences (the base beliefs) and a set of inference rules. An agent's belief set is the set of all sentences that can be derived from the base beliefs by exhaustive application of the inference rules. Agents that are not logically omniscient can exist in the deduction model, because the set of inference rules is not required to be complete.

In [13] we give a detailed comparison of the computational model and the deduction model. To summarize a few of the main points of that comparison,

- There are belief machines for which no equivalent set of inference rules exists. Constraints C1–C5 rule out many of them, but not all. In particular, the constraints allow machines that refuse to believe certain combinations of sentences. For example, there exist belief machines that satisfy C1–C5 and for which $B(a, P(c)) \wedge B(a, \neg P(c))$ is unsatisfiable. In the deduction model, any set of sentences can be believed simultaneously, because the model doesn't constrain the set of base beliefs in any way.

- There are deduction structures for which no equivalent belief machine exists, namely deduction structures with inference rules whose closure is not computable—for example, a complete set of inference rules for FOL. Obviously, prohibiting this kind of idealized agent was an explicit goal in the design of our model.

- Konolige proves completeness for his logic, while we have shown that ours is incomplete. The difference comes from the fact that Konolige allows the a deduction structure to have an infinite set of base beliefs. This is no advantage in terms of modeling real agents, since the base beliefs are intended to model the facts that an agent has explicitly learned, and this set is necessarily finite for any real agent.

- Our model is easier to use as a tool for understanding simulative inference in practical systems, since most real reasoning systems do not work by computing the closure of a set of inference rules, but rather involve complex control algorithms.

- There are also a few differences that stem not from the difference between belief machines and deduction structures, but from other choices in the designs of the two logics:

  ID constants: Konolige requires that for each agent, there be a naming map that maps each individual in the domain to a unique canonical name, called an *id constant*. ID constants are used in the semantics of quantifying-in: an open belief formula $\exists x B(a, P(x))$ (translating Konolige's notation into ours) is true iff there is some id constant κ for which B(a,P(κ)) is true. We have chosen not to require that individuals have canonical names, and we have defined the semantics so that a quantified-in formula $\exists x B(a, P(x))$ is true if there is any term τ for which $B(a, P(\tau))$ is true.

  Quantification over believers: In Konolige's logic, there is a different belief operator for each agent, rather than a single operator that takes a believer argument. Therefore, quantification over believers is impossible in Konolige's logic—there is no equivalent of the formula $\forall x B(x, P(c))$ of our logic.

  Equality: Our logic includes an equality operator, while Konolige's does not. This is one of the things that makes the (restricted) completeness proof for our logic more complex.

4 Conclusion

We have introduced a new logical semantics of belief in which reasoning is modeled as the application of a reasoning algorithm. The model serves as a tool for exploring the technique of simulative inference, that is, reasoning about another agent's beliefs by simulating its thought process. We do not specify a particular reasoning algorithm, but rather define a family of logics, where each possible reasoning algorithm gives rise to a different logic. We have defined two inference rules for reasoning about an agent's beliefs, and we have shown that the rules are sound if the reasoning algorithm satisfies certain input-output constraints. The logic is not complete, but a syntactically-restricted variant of it is complete.

For the sake of brevity, definitions and proofs were summarized or omitted entirely in this paper. For details, see [13]. That reference also describes a case study in which our logic was used to add simulative reasoning about belief to an existing logical inference system, called EPILOG.

5 Acknowledgments and prior publications

This paper is a summary of my Ph.D. work, which was performed at the University of Rochester under the supervision of Lenhart Schubert, with the support of NSF research grants numbers IRI-9623665 and IRI-9503312, and U.S. Air Force/Rome Labs research grant number F30602-97-1-0348. The most recent and comprehensive publication is [13]. Versions of the computational model of belief had previously been introduced in [15] and [16]; introspection and nonmonotonic belief machines were explored in [12]; issues relating to the efficient implementation of simulative inference were explored in [14].

References

[1] Thomas Ågotnes and Natasha Alechina. Semantics for dynamic syntactic epistemic logics. In *Proceedings of the 10th International Conference on Principles of Knoweledge Representation and Reasoning*, 2006.

[2] James P. Delgrande. A framework for logics of explicit belief. *Computational Intelligence*, 11(1):47–88, 1995.

[3] Ho Ngoc Duc. *Resource-Bounded Reasoning about Knowledge*. PhD thesis, Leipzig University, 2001.

[4] Rolf A. Eberle. A logic of believing, knowing, and inferring. *Synthese*, 26:356–382, 1974.

[5] Jennifer Elgot-Drapkin, Sarit Kraus, Michael Miller, Madhura Nirkhe, and Donald Perlis. Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-4072, University of Maryland Computer Science Department, College Park, MD, October 1999.

[6] Ronald Fagin and Joseph Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[7] Adam J. Grove. Naming and identity in epistemic logic part II: a first-order logic for naming. *Artificial Intelligence*, pages 311–350, 1995.

[8] Andrew R. Haas. A syntactic theory of belief and action. *Artificial Intelligence*, 28:245–292, 1986.

[9] Andrew R. Haas. An epistemic logic with quantification over names. *Computational Intelligence*, 1995.

[10] Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. Algorithmic knowledge. In Ronald Fagin, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pages 255–266, San Francisco, 1994. Morgan Kaufmann.

[11] Jaakko Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, New York, 1962.

[12] Aaron N. Kaplan. Simulative inference about nonmonotonic reasoners. In *Theoretical Aspects of Rationality and Knowledge: Proceedings of the Seventh Conference*, pages 71–81. Morgan Kaufmann, 1998.

[13] Aaron N. Kaplan. *A Computational Model of Belief*. PhD thesis, University of Rochester, 2000.

[14] Aaron N. Kaplan. Reason maintenance in a hybrid reasoning system. *Journal of Language and Computation*, 1(2):227–240, 2000. Preliminary version presented at Workshop on Inference in Computational Semantics, Amsterdam, August 1999.

[15] Aaron N. Kaplan and Lenhart K. Schubert. Simulative inference in a computational model of belief. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning*, Studies in Linguistics and Philosophy, pages 185–202. Kluwer, 1999. Preliminary version presented at International Workshop on Computational Semantics, Tilburg, The Netherlands, January 1997.

[16] Aaron N. Kaplan and Lenhart K. Schubert. A computational model of belief. *Artificial Intelligence*, 120:119–160, 2000.

[17] Kurt Konolige. *A Deduction Model of Belief*. Morgan Kaufmann Publishers, Inc., Los Altos, California, 1986.

[18] S. A. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

[19] Gerhard Lakemeyer. Limited reasoning in first-order knowledge bases. *Artificial Intelligence*, 71:213–255, 1994.

[20] Hector. J. Levesque. A logic of implicit and explicit belief. In *Proceedings AAAI-84*, pages 198–202, Austin, 1984.

[21] Donald Perlis. Languages with self-reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34:179–212, 1988.

[22] Jan Plaza. Logics of public communications. In *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216. North-Holland, 1989.

[23] Riccardo Pucella. Deductive algorithmic knowledge. *Journal of Logic and Computation*, 16(2):287–309, 2006.

[24] H.P. van Ditmarsch, W. van der Hoek, and B.P. Kooi. Concurrent dynamic epistemic logic. In V.F. Hendricks, K.F. Jorgensen, and S.A. Pedersen, editors, *Knowledge Contributors*, volume 322 of *Sythese Library*. Kluwer Academic Publishers, Dordrecht, 2003.

Diversity of Agents

Fenrong Liu *

May 10, 2006

Abstract

Diversity of agents is investigated in the context of standard epistemic logic, dynamic information update, and belief revision. We provide a systematic discussion of different sources of diversities, such as introspection ability, powers of observation, memory capacity, and revision policies. In each case, we show how this diversity can be encoded in a logical system allowing for individual variation among rational agents. We conclude by raising some general issues concerning this view of a logic as a system for encoding a society of diverse agents and their interaction.

1 Diversity Inside Logical Systems

Logical systems seem to prescribe one norm for an "idealized agent". Any discrepancies with actual human behavior are then irrelevant, since the logic is meant to be normative, not descriptive. But logical systems would not be of much appeal if they did not have a plausible link with reality. And this is not just a matter of confronting one ideal norm with one kind of practical behavior. The striking fact is that human and virtual agents are not all the same: actual reasoning takes place in societies of diverse agents.

This diversity shows itself particularly clearly in *epistemic logic*. There have been long debates about the appropriateness of various basic axioms, and they have to do with agents' different powers. In particular, the modal Distribution Axiom has the following epistemic flavor:

Example 1.1 Logical omniscience: $K(\varphi \to \psi) \to (K\varphi \to K\psi)$.

Do rational agents always *know the consequences* of what they know? Most philosophers deny this. There have been many attempts at bringing the resulting diversity into the logic as a legitimate feature of agents. Some authors have used "awareness" as a sort of restriction on short-term memory ([FH85]), others have concentrated on the stepwise dynamics of making inferences ([Kon88], [Dun95]). A well-informed up-to-date philosophical summary is found in [Egr04].

The next case for diversity lies in a different power of agents:

Example 1.2 Introspection axioms: $K\varphi \to KK\varphi$, $\neg K\varphi \to K\neg K\varphi$.

Do agents *know when they know* (or *do not know*)? Many philosophers doubt this, too. This time, there is a well-established way of incorporating different powers into the logic, using different accessibility relations between possible worlds in Kripke models. Accordingly, we get different modal logics: K, T, $S4$, or $S5$. Each of these modal logics can be thought of as describing one sort of agents. The interesting setting is then one of combinations. E.g., a combined language with two modalities K_1, K_2 describes a two-person society of introspectively different agents! This gives an interestingly different take on current logic combinations ([GS98], [KZ03]): the various ways of forming combined logics, by "fusions" $S5 + S4$ or "products" $S5 \times S4$, correspond to different assumptions about how the agents *interact*. Effects may be surprising here. E.g., later on, in our discussion of memory-free agents, we see that knowledge of memory-free agents behaves much like "universal modalities". But in certain modal logic combinations,

*I would like to thank Johan van Benthem and three anonymous reviewers for their constructive comments that greatly improved this paper.

adding a universal modality drives up complexity, showing how the interplay of more clever and more stupid agents may itself be very complex...

Thus, we have seen how *diversity exists inside standard epistemic logic*, and hence likewise in doxastic logic. The purpose of this paper is to bring to light some further sources of diversity in existing logics of information. Eventually, we would want to move from complaints about 'limitations" and "bounds" to a positive understanding of how societies of diverse agents can perform difficult tasks ([GTtARG99]). But our actual contribution is more modest, viz. a discussion of sources of diversity in dynamic logics of information. Section 2 is a brief identification of further parameters of variation for agents beyond those of standard epistemic logic. Section 3 looks at dynamic epistemic logics of information update, showing how limited powers of observation are accounted for, and adding some new systems with bounded memory. Section 4 takes a parallel look at dynamic doxastic logics for belief revision, and shows how different revision policies can be dealt with. Finally, Section 5 is a brief summary, which also identifies some further more ambitious questions.

This paper is based on existing literature plus unpublished work in the author's Master's Thesis ([Liu04]). We will mainly cite technical results, and put them into a hopefully fresh story.

2 Sources of Diversity

The diversity of logical agents seems to stem from different sources. In what follows, we shall mainly speak about "limitations", even though this is a loaded term suggesting "failure". The more cheerful reality is of course that agents have various resources, and they use these positively to perform tasks, often highly successfully.

Our epistemic axioms point at several "parameters" of variation of agents:

(a) *inferential/computational power*: making all possible proof steps,

(b) *introspection*: being able to view yourself in "meta-mode".

One further potential parameter relevant to epistemic logic is the "awareness" studied by some authors([FH85]), which suggests some resource like limited attention span, or short-term memory.

Next, consider modern dynamic logics of information, whose motivation sounds closer to actual cognitive practice. These also turn out to incorporate idealizations that suggest further parametrization for diversity. We start with the case of information update.

Consider the basics of public announcements logic (PAL): $!\varphi$ in the language is intended to mean "a fact φ is truthfully announced". PAL considers the epistemic effects those actions can bring about. In addition to static axioms that "invite diversity", here is one more. The following principle is crucial to the way PAL analyzes epistemic effects of assertions:

$[!\varphi]K_i\psi \leftrightarrow \varphi \rightarrow K_i[!\varphi]\psi$ *Knowledge Prediction Axiom*

But the validity of this axiom presupposes several things, notably *Perfect Observation* and *Perfect Recall* by agents. The event of announcement must be clearly identifiable by all, and moreover, the update induced by the announcement only works well on a unique current information state recording all information received so far. More technically, these points show in a detailed soundness proof for the Knowledge Prediction Axiom in its intended semantics. We will discuss this in Section 3, in the more general framework of "product update" for dynamic epistemic languages ([BMS98]). Thus, we have found two more parameters of diversity in logic. Agents can differ regarding:

(c) *observation*: stipulate agents' powers of observation for current events,

(d) *memory*: stipulate agents' limited memory capacity, e.g., store only the last k events observed, for some fixed k.

Can one deal with this inside the logic? As we will see, dynamic epistemic logic with product update can itself be viewed as a calculus of observational powers. And as to memory, [BL04] have shown how to incorporate this into dynamic epistemic logic (DEL) for memory-free agents, and we will extend their style of analysis below to arbitrary finite memory bounds.

Yet another source for diversity of agents lies in *belief revision theory* ([AGM85]). This time, agents must revise their beliefs on the basis of incoming information which may contradict what they believed so far. This scenario is different from the preceding one, as has been pointed out from the start in this area ([GR95]). Even for agents without limitations of the earlier sorts, there is now another legitimate source of diversity, viz. their habits that create diversity:

(e) *revision policies*: varying from conservative to radical revision.

Different agents may react differently towards new information: some behave conservatively and try to keep their original beliefs as much as possible, others may be radical, easily accepting new information without much deliberation. However, these policies are not explicitly part of belief revision theory, except for some later manifestations ([Was00]). We will show in this paper, following [Liu04], [BL06], how they can be brought explicitly into dynamic logic as well.

This concludes the list of parameters of diversity that we see in current dynamic-epistemic and dynamic-doxastic logics. It is important to mention that acknowledging this diversity inside logical systems is not a concession to the ugliness of reality. It is rather an attempt to get to grips with the most striking aspect of human cognition: despite our differences and limitations, societies of agents like us manage to cooperate in highly successful ways! Logic should not ignore this, but rather model it and help explain it. Our paper is a modest attempt at systematization toward this goal.

3 Dynamic Logics of Information Update

3.1 Preliminaries: Product Update

To model knowledge change due to incoming information, the powerful mechanism is dynamic epistemic logic, which has been developed intensively by [Pla89], [Ben96], [BMS98], [Ger99], [DHK06], etc. Since our discussions in this paper will be heavily based on DEL, we briefly recall its basic ideas and techniques.

Definition 3.1 [(epistemic model)] An epistemic model is a tuple $\mathcal{M} = (S, \{\sim_i | i \in G\}, V)$ [1] such that S is a non-empty set of states, G is a group of agents, each $\sim_i$ is a binary epistemic relation on S, and V is a function assigning to each proposition variable p in Φ a subset $V(p)$ of S. $\lhd$

Definition 3.2 [(event model)] An event model is a tuple $\mathcal{E} = (E, \sim_i, PRE)$ such that E is a non-empty set of events, $\sim_i$ is a binary epistemic relation on E, PRE is a function from E to the collection of epistemic propositions. $\lhd$

Note that we have a new function PRE in a event model, the intuition is that it gives the *preconditions* for an action: an event e can be performed at world s only if the world s fulfills the precondition $PRE(e)$.

Definition 3.3 [(product update)] Let an epistemic model $\mathcal{M}$=$(S, \sim_i, V)$ and an event model $\mathcal{E}$=$(E, \sim_i, PRE)$ be given. The product update model is defined to be the model $\mathcal{M} \otimes \mathcal{E}$=$(S \otimes E, \sim'_i, V')$:

- $S \otimes E = \{(s, e) \in S \times E$: $(\mathcal{M}, s) \models PRE(e)\}$
- $(s, e) \sim'_i (t, f)$ iff both $s \sim_i t$ and $e \sim_i f$
- $V'(p) = \{(s, e) \in \mathcal{M} \otimes \mathcal{E}$: $s \in V(p)\}$.

$\lhd$

[1] I will sloppily write it as $\mathcal{M} = (S, \sim_i, V)$ when G is clear from the context.

The *actual world* of the new model is the pair consisting of the actual world in $\mathcal{M}$ and the actual event or action in $\mathcal{E}$. The product rule says that uncertainty among new states can only come from existing uncertainty via indistinguishable actions. The above notions suggest an extension of the epistemic language.

Definition 3.4 [(dynamic epistemic language)] Let a finite set of proposition variables Φ, a finite set of agents G, a finite set of events E be given. The dynamic epistemic language is defined by the rule

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid [\mathcal{E}, e]\varphi$$

where $p \in \Phi$, $i \in G$, and $e \in E$. $\lhd$

Normally, one could also add the usual action operations of composition, choice, and iteration from propositional dynamic logic to the event vocabulary. The language has new dynamic modalities $[\mathcal{E}, e]$ referring to epistemic events, and these are interpreted in the product update model as follows:

$$\mathcal{M}, s \models [\mathcal{E}, e]\varphi \text{ iff } \mathcal{M}\otimes\mathcal{E}, (s, e) \models \varphi.$$

Reduction axioms in DEL play an important role in encoding the epistemic changes. For example, the following axiom concerns agents' knowledge change.

$$[\mathcal{E}, e]K_i\varphi \leftrightarrow PRE(e) \rightarrow \bigwedge_{f\in\mathcal{E}}\{K_i[\mathcal{E}, f]\varphi : e \sim_i f\}.$$

Intuitively, after an event e takes place the agent i knows φ, is equivalent to saying that if the event e can take place, i knows beforehand that after e (or any other event f which i can not distinguish from e) happens φ will hold. Such a principle is of importance in that it allows us to relate our knowledge after an action takes place to our knowledge beforehand, which plays a crucial role in communication, and planning in general.

PAL is the simplest case of update logic, in the sense that the event model contains one single event. Moreover, the precondition of the action $!\varphi$ boils down to the fact that φ is true, as we will see in the formulas in the next section. In this paper, for easy understanding, we use PAL to motivate our claims, though we also consider things within DEL with a general mechanism of product update.

3.2 Public Announcement, Observation and Memory

First, we recall the complete axiom system for public announcement.

Theorem 3.5 ([*Pla89*][*Ger99*]) *PAL is axiomatized completely by the usual laws of epistemic logic plus the following reduction axioms:*

($!p$). $[!\varphi]p \leftrightarrow \varphi \rightarrow p$ *for atomic facts p*

($!\neg$). $[!\varphi]\neg\psi \leftrightarrow \varphi \rightarrow \neg[!\varphi]\psi$

($!\wedge$). $[!\varphi]\psi \wedge \chi \leftrightarrow [!\varphi]\psi \wedge [!\varphi]\chi$

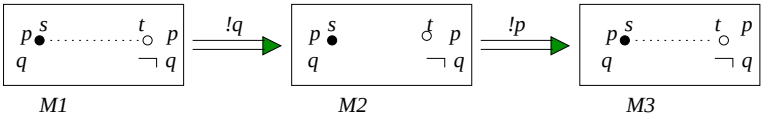
($!K$). $[!\varphi]K_i\psi \leftrightarrow \varphi \rightarrow K_i[!\varphi]\psi$

Next, to introduce variety in *observation*, we need to assume a set of possible announcements $!\varphi$, $!\psi$, ... where agent i need not be able to distinguish all of them. This uncertainty can be modelled by an equivalence relation $\sim_i$ between statements which i cannot distinguish. The following principle may then be proved:

Theorem 3.6 ([*BL04*]) *The following reduction axiom is valid for agents with limited powers of observation:*

$$[!\varphi]K_i\chi \leftrightarrow (\varphi \rightarrow K_i \textstyle\bigwedge_{!\psi\sim_i!\varphi}[!\psi]\chi)$$

As we have seen from the previous section, *Perfect Recall* assumes that agents can remember all the events happened so far. But in reality there are agents with bounded memory, who can only remember a fixed number of previous events. It is much harder in PAL to model memory difference because the world elimination update procedure shifts agents to ever more informed states. How can they forget? Here is one option (suggested by [BL04]). First, we can reformulate PAL semantics as in [BL06] to never eliminate worlds. The idea is to let announcement $!\varphi$ cut all links between φ-worlds and $\neg\varphi$-worlds, but keep all worlds in. Now, the resulting "unreachabilities" represent the information agents have so far. One way of describing a memory-restricted agent is then as having forgotten part or all of the "missing links". In the most extreme case, a *memory-free* agent will only acknowledge distinctions made by the last announcement. Apart from that, worlds will be indistinguishable. Agents like this do not satisfy the earlier reduction axiom $(!K)$, see the following example.



Example 3.7 There are two possible worlds, s and t in $\mathcal{M}_1$, p and q hold at s, p and $\neg q$ hold at t. After q is announced, we get a new model $\mathcal{M}_2$, in which there is no uncertainty link between s and t. Then we have $(\mathcal{M}_2, s) \models p \to K_i(p \to q)$, i.e. $(\mathcal{M}_2, s) \models p \to K_i[!p]q$. After that, p is announced, and we have $\mathcal{M}_3 \not\models K_i q$, since the agent forgot $!q$ already. We look back at $\mathcal{M}_2$: $(\mathcal{M}_2, s) \not\models [!p]K_i q$. The reduction axiom does not hold!

Fact 3.8 *The correct axiom valid for memory-free agents is*

$$[!\varphi]K_i\psi \leftrightarrow (\varphi \to U[!\varphi]\psi)$$

With the above picture, it is easy to check that the axiom is correct. Here $U\varphi$ is an *universal modality* saying that φ holds in all worlds. To restore the harmony of an update logic, one should also extend the update reduction axioms with a clause for the new operator U. The following one is valid:

$$[!\varphi]U\psi \leftrightarrow \varphi \to U[!\varphi]\psi$$

Note that it looks like a $(!K)$ clause.

Thus, "logic of public announcement" is actually a family of formal systems, depending on the chosen update rule, which in turn depends on the memory type of the agents.

3.3 Adding Memory to Product Update

We now extend the update mechanism to agents with finite memory. As we have seen above, for memory-limited agents, the main point is to try to keep all the possible worlds around, so that the agent can always refer to the possible worlds which have been deleted before. We are still working with the *DEL* models, where information is changed by events. By a *k-memory* agent, we mean an agent that remembers only the last k events before the most recent one. A 0-memory or memory-free agent, then, knows only what she learned from the most recent action; an agent with memory of length 1 knows only what she learned from the most recent action and the one before it, and so on. Now we must define the corresponding updates:

Definition 3.9 ([Sny04]) Let $\mathcal{M}$ be an epistemic model, $\mathcal{E}_{-k}$ be the k-th event model before the most recent one $\mathcal{A}$. The product update for k-memory agents is defined as

(1) $\mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1} \times \mathcal{E}$
$=\{(s, a_{-k}, \ldots, a_{-1}, a) : (s, a_{-k}, \ldots, a_{-1}) \in \mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1}$ and $a \in \mathcal{E}\}$

(2) $(s, a_{-k}, \ldots, a_{-1}, a) \sim_i (t, b_{-k}, \ldots, b_{-1}, b)$ iff $(\mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1} \models PRE(a)$ iff $\mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1} \models PRE(b))$ and $a \sim_i b$.

$\lhd$

Compare with the standard product update definition, in the above definition (1) leaves the precondition restriction out. This simply makes it possible to keep all the worlds around. (2) gives restrictions to the states, and defines the uncertainty relation in the new models.

Another alternative is to introduce an auxiliary *copy action* $C!$ which takes place everywhere. It puts those worlds which are to be deleted into a stack, and makes sure agents can always find them when needed.

Definition 3.10 ([Liu04]) Let $\mathcal{M}$ be an epistemic model, $\mathcal{E}_{-k}$ be the k-th action model before the most recent one $\mathcal{E}$. The product update for k-memory agents is defined as

$(1')$ $\mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1} \times \mathcal{E}$
$=\{(s,a_{-k},\ldots,a_{-1},a) : (s,a_{-k},\ldots,a_{-1}) \in \mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1}$ and $a \in \mathcal{E}$ and $(s,a_{-k},\ldots,a_{-1}) \models PRE(a)\}$

$(2')$ For $a_{-k},\ldots,a_{-1},a,b_{-k},\ldots,b_{-1},b \neq C!$,
$(s,a_{-k},\ldots,a_{-1},a) \sim_i (t,b_{-k},\ldots,b_{-1},b)$ iff $a_{-k} \sim_i b_{-k},\ldots,a_{-1} \sim_i b_{-1}$ and $a \sim_i b$

$\lhd$

These two definitions can take care of our goal, namely, to keep those worlds around in the model. The technical difference lies in their different intuitions. In Def 3.9, it is believed that all the possible worlds make sense for bounded memory agents, so one *should not* remove the worlds because of the precondition restriction. Differently, Def 3.10 makes the worlds stay around with the help of the auxiliary copy action, just as memory bounded agents often do in real life.

3.4 Discussion

We have identified two new parameters of variation for dynamic updating agents; powers of observation, and powers of memory. DEL as it stands provides a way of modelling the former, while we have shown how it can also be modified to accommodate update for agents with bounded memory.

Of course, this is only the beginning of an array of further questions. In particular, we would like to have a more structured account of memory, as in computer science where we update data or knowledge bases. Update mechanisms are more refined there, referring to memory structure with actions such as information replacement([Liu04]). This is one instance of a more "constructive" syntactic approach to update, complementary to our abstract one in terms of model manipulation. Whether our current semantic method or a syntactic one works better for finding agents' parameters of diversity seems a question worth investigating.

Other questions that come up have to do with the interaction between agents. Our logical systems can describe the behavior of various agents via reduction axioms, that is, schemas with infinitely many concrete instances. But they cannot yet state in one single formula "that an agent is of a certain type". And as they stand, they are even less equipped to describe the interplay of different agents in a compact illuminating way. Thus, we have only established a first toehold for diversity within the bastion of dynamic epistemic logic.

4 Diversity in Dynamic Logics of Belief Change

Belief revision describes what happens when an agent is confronted with new information which conflicts her earlier belief. Different policies toward revising beliefs fall within the AGM postulates. We first look at a concrete example of how belief revision can be implemented technically.

4.1 Belief Revision as Relation Change

Example 4.1 ([Ben06]) ($\Uparrow$) Radical revision
$\Uparrow P$ is an instruction for replacing the current ordering relation $\leq$ between worlds by the following: all P-worlds become better than all $\neg P$-worlds, and within those two zones, the old ordering remains.

Another possibility would be that just the best P-world comes to the top, leaving the further order unchanged. A more general description of such policies can be given as ways of *changing a current preference relation* ([BL06], [Rot06]). Viewed in this way, the dynamic logic for radical revision can be axiomatized completely in DEL style:

Theorem 4.2 ([*Ben06*]) *The dynamic logic for radical revision* ($\Uparrow$) *is axiomatized completely by a complete axiom system KD45 on the static models, plus the following reduction axioms*

($\Uparrow p$). $[\Uparrow \varphi]p \leftrightarrow p$

($\Uparrow \neg$). $[\Uparrow \varphi]\neg\psi \leftrightarrow \neg[\Uparrow \varphi]\psi$

($\Uparrow \wedge$). $[\Uparrow \varphi](\psi \wedge \chi) \leftrightarrow [\Uparrow \varphi]\psi \wedge [\Uparrow \varphi]\chi$

($\Uparrow B$). $[\Uparrow \varphi]B_i\psi \leftrightarrow (E\varphi \wedge B_i([\Uparrow \varphi]\psi|\varphi)) \vee B_i[\Uparrow \varphi]\psi$[2]

The last axiom here shows the doxastic effects of the chosen policy. But it still does so somewhat implicitly. In what follows, we will explore the same issues, but now with a notion of *plausibility* for worlds, which allows us to high-light policies more directly.

4.2 Plausibility Change

We first review briefly some previous work in this line. Following [Spo88], a κ-ranking function was introduced into the dynamic epistemic logic in [Auc03]. A κ-ranking is a function κ from a given set S of possible worlds into the class of numbers up to some maximum Max. The numbers can be thought of as denoting degree of surprise. 0 denotes 'unsurprising', 1 denotes 'somewhat surprising', etc. In other words, κ represents a plausibility grading of the possible worlds. This makes it possible to express the degree of beliefs.

Next, we also add plausibilities κ^* to the event model $\mathcal{E}$, representing the agents' view on which event is taking place. With plausibilities assigned to states and events, belief changes via product update. Here is the key formula:

$$\kappa_i'(s,e) = Cut_{Max}(\kappa_i(s) + \kappa_i^*(e) - \kappa_i^s(\varphi)),$$

where $\varphi = PRE(e)$, $\kappa_i^s(\varphi) = min\{\kappa_i(t) : t \in V(\varphi) \text{ and } t \sim_i s\}$

$$Cut_{Max}(x) = \begin{cases} x & \text{if } 0 \leq x \leq Max \\ Max & \text{if } x > Max. \end{cases}$$

The crucial reduction axiom for belief in [Auc03] is the following:

$[\sigma_j, \psi]B_i^m\varphi \leftrightarrow (\psi_j \rightarrow \bigwedge\{B_i^{l-1}\neg\psi_k \wedge \neg B_i^l\neg\psi_k \rightarrow B_i^{m+l-\kappa_i^*(\sigma_k)}$
$[\sigma_k, \psi]\varphi : \sigma_k \sim_i \sigma_j$, and $l \in \{0, \ldots, Max\}\})$ where $m < Max$

Here σ_j and σ_k are actions, ψ are preconditions. $B_i^m\varphi$ is intended to mean that an agent believes φ up to degree m, i.e. φ is true in all epistemically accessible worlds of κ-value $\leq m$.

In the following section, instead, we take a more perspicuous approach, using epistemic-doxastic language with propositional constants to describe the plausibility change.

4.3 Revision Policies

What follows is taken from the unpublished Master's thesis [Liu04].

Definition 4.3 The *epistemic-doxastic language* is defined as

[2] Some explanations about notations here: E is the existial modality, the dual of the universal modality U. The symbol | is to denote an conditionalization and it is intended to mean 'given that'.

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid q_i^\alpha$$

where $p \in \Phi$, a set of propositions, $i \in G$, a set of agents, and α is a κ-value in $\mathbb{N}$, q_i^α are a special type of propositional constants. $\lhd$

The interpretation is as usual, but with the following *truth condition* for the additional propositional constants:

$$(\mathcal{M}, s) \models q_i^\alpha \text{ iff } \kappa_i(s) \leq \alpha$$

The update mechanism can now be defined by merely specifying the new κ-value in a product model $\mathcal{M} \times \mathcal{E}$. To keep our discussion simple, we use just this:

Definition 4.4 [(Bare addition rule)] The new plausibilities in product models are defined by the following rule:

$$\kappa_i'(s, e) = \kappa_i(s) + \kappa_i^*(e)$$

$\lhd$

Theorem 4.5 ([*Liu04*]) *The complete dynamic logic of plausibility belief revision consists of the key reduction axioms in Theorem.1 plus the following new one.*

$$[\varphi!]q_i^\alpha \leftrightarrow q_i^{\alpha - \kappa_i(\varphi!)}$$

Proof. This Axiom captures the plausibility change. By the Bare addition rule, the plausibility of the world in the original model $\kappa_i(w) = \kappa_i(w, e) - \kappa_i(e) = \alpha - \kappa_i(\varphi!)$. ■

In fact, more generally, the plausibility rule can be any function of the plausibility of the previous event and that of the previous state. This is of course the locus for different policies! Moreover, if the update rule is functionally expressible, we can still get a complete logic, though clearly the simple substraction will not work anymore. To illustrate how it works, we now present a proposal which attempts to incorporate more elements into the update rule to characterize the diverse aspects of agents.

Definition 4.6 ([Liu04]) Let the weight that an agent i gives to the state s be λ and the weight to the event e be μ. The plausibility of the new state (s, e) is calculated by the following rule

$$\kappa_i'(s, e) = \frac{1}{\lambda + \mu}(\lambda\kappa_i(s) + \mu\kappa_i^*(e)) \quad (\natural).$$

$\lhd$

The variations of parameter λ and μ describe a range of various agents. For instance, when μ=0, we get *highly conservative agents*, the ($\natural$) rule turns into $\kappa_i'(s, e) = \kappa_i(s)$. It means the agent does not consider the effect of the action. Similarly, λ=0, the agents are *highly radical*, and $\kappa_i'(s, e) = \kappa_i^*(e)$. When $\lambda = \mu$, we call them *Middle of the road agents* who believe plausibility of states and actions should play an equally important role in determining the plausibility of the new state. In this manner, we have distinguished five types of agents. For an even more general view of agents' behavior towards incoming information, see the continuing work in [Liu06].

In particular, this view of policies challenges one key assumption of AGM: the Success Postulate, which accords top status to the new information. Highly conservative agents would not take new information, right from the beginning, the belief revision cannot go on successfully. One can also get complete dynamic logics for various policies, but we will not pursue these here.

4.4 Comparisons and Discussion

Our treatment of belief revision provides a simple format of plausibility change, where different policies show in a perspicuous manner in the reduction axiom for the "value constants". Moreover, our treatment also goes beyond the standard AGM paradigm, in that we allow agents to doubt the current information. Here are a few further issues that come up in this setting, some conceptual, some technical.

First, doubting the current information might also make sense for PAL and DEL scenarios without belief revision. It is easy to achieve this by adding further events to an event model, providing, say, a public announcement $!\varphi$ with a counterpart $!\neg\varphi$ with some plausibility value reflecting the strength of the "dissenting voice". Likewise, policies with weights for various factors in update make much sense in recently proposed dynamic logics of probabilistic update ([Auc05], [BGK06]).

Incidentally, this DEL approach via modified event models for different policies may also suggest that we can *relocate* policies from "modified update rules" to "modified event models" with a standard update rule. We must leave this comparison to another occasion.

Next, there is an issue about relation-changing views of belief revision as in Section 4.1 versus our plausibility changes. One obvious difference is that plausibility change stays within the realm of *connected* world-ordering relations, whereas relational redefinition need not. On the other hand, plausibility change can describe scenarios such as "add one plausibility point to every φ-world", which have no immediate counterpart in terms of relation changes. For comparison, we refer to [Liu06]. Of course, all general questions from Section 3.4 about representing agents and their *types* return here in even greater force.

Finally, a new observation concerning two parameters, revision policy and memory: Technically speaking, the update behavior of highly radical agents is no different from that of memory-free agents, as they simply take the new information without considering what happened before (of course for different reasons). In other words, the event that takes place completely characterizes the "next" epistemic state of the agent. That seems to be related also to notions such as "only knowing" or "minimal knowledge" in [Lev90] and [HJT90]. This observation seems to suggest a way to unify some of our parameters discussed so far.

5 Conclusion: A Unified Account of Diversity?

We have investigated many different sources of diversity, some visible in static logics, some in dynamic ones. Besides the old parameters from epistemic logic, namely computation and introspection ability, we have added several new aspects, i.e. observation power, memory capacity and revision policy. Our discussion has been mostly in the framework of dynamic epistemic logic and we have shown how it is possible to allow for a characterization of diversity within the logic. To summarize, look at the following diagram consisting of the main components of dynamic epistemic logic:

$$\begin{cases} Static \quad language & \text{Epistemic model } \mathcal{M}; \\ Dynamic \quad language & \text{Event model } \mathcal{E}; \\ Product \quad update & \text{Model change } \mathcal{E} \times \mathcal{M}. \end{cases}$$

In the previous sections we have shown that the diversity of agents can be explicitly modelled in terms of these logical components. The following table is an outline of our discussions.

Component	Residence	Diversity
$\mathcal{M}$	relations between worlds	introspection
$\mathcal{E}$	relations between actions	observation
$\mathcal{M} \times \mathcal{E}$	update mechanism	memory, revision policy

As we can see from the table, by introducing parameters of variation to each component, we are able to describe diversity of agents inside the logic system. Note that computation ability is not included in the table, we think its dimension is slightly different.

Still, there remains the issue whether one can have a *general* view of the natural "parameters" that determine differences in behavior of logical agents. Our analysis does not provide such a general account, but at least, it shows more richness and uniformity than earlier ones. Especially, we have seen one possibility to unify the parameters of revision policy and memory capacity at the end of the previous section. If more uniformity is needed, a good challenge would be to unify our observation- and memory-based analysis with diversity in deductive and computational powers. Our current speculation would be that many "idealizations" in standard logic have something to do with passing to a *countable limit*. Closing knowledge under consequence, computes the k-step consequences for all successive k. Introspection involves computing the transitive closure of the base accessibility relation. And memory goes from ever larger k-memory agents to unbounded stacks. At some abstract level, this may all be computing some fixed-point for a closure operator in some superstructure over an existing logic.

Even so, we hope that our account of diversity provides a fresh look at "logical systems". We now see them as vehicles for agents having powers of observation, memory, inference, computation, attention, and so on. Indeed, they begin to look remarkably like us!

References

[AGM85] C. Alchourrón, P. Gardenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Auc03] G. Aucher. A combination system for update logic and belief revision. Master's thesis, ILLC, University of Amsterdam, 2003.

[Auc05] G. Aucher. How our beliefs contribute to interpret actions. In L. Z. Varga M. Pechoucek, P. Petta, editor, *CEEMAS 2005*, pages 276 – 286. Springer, 2005. LNAI 3690.

[Ben96] J. van Benthem. *Exploring Logical Dynamics*. CSLI Publication, Stanford, 1996.

[Ben06] J. van Benthem. Dynamic logic for belief change. ILLC Research Reports, PP-2006-11, University of Amsterdam, 2006.

[BGK06] J. van Benthem, J. Gerbrandy, and B. Kooi. Dynamic update with probabilities. ILLC Research Reports, PP-2006-21, Universtity of Amsterdam, 2006.

[BL04] J. van Benthem and F. Liu. Diversity of logical agents in games. *Philosophia Scientiae*, 8(2):163–178, 2004.

[BL06] J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic*, 2006. To appear.

[BMS98] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge (TARK 98)*, pages 43–56, 1998.

[DHK06] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, Berlin, 2006. To appear.

[Dun95] P.H. Dung. An argumentation-theoretic foundation for logic programming. *Journal of Logic Programming*, 22:151–177, 1995.

[Egr04] P. Egre. *Propositional Attitudes and Epistemic Paradoxes*. PhD thesis, Universite Paris 1 et IHPST, 2004.

[FH85] R. Fagin and J. Y. Halpern. Belief, awareness, and limited reasoning. In *In Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, 1985.

[Ger99] J. Gerbrandy. *Bisimulation on Planet Kripke*. PhD thesis, ILLC Amsterdam, 1999.

[GR95] P. Gardenfors and H. Rott. Belief revision. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4. Oxford University Press, 1995.

[GS98] D. Gabbay and V. Shehtman. Products of modal logics. *Logic Journal of the IGPL*, 6(1):73–146, 1998. part 1.

[GTtARG99] G. Gigerenzer, P. Todd, and the ABC Research Group. *Simple Heuristics that Make us Smart*. New York: Oxford University Press, 1999.

[HJT90] W. van der Hoek, J. Jaspars, and E. Thijsse. A general approach to multi-agent minimal knowledge. In G. Brewka M. Ojeda-Aciego, I.P. Guzman and L.M. Pereira, editors, *Proceedings of the 7th European Workshop on Logics in Artificial Intelligence (JELIA 2000)*, pages 254–268. Springer-Verlag, Heidelberg, 1990. LNAI 1919.

[Kon88] K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35(3):343–382, 1988.

[KZ03] A. Kurucz and M. Zakharyaschev. A note on relativised products of modal logics. In F. Wolter P. Balbiani, N.-Y. Suzuki and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 4, pages 221–242. King's College Publications, 2003.

[Lev90] H. J. Levesque. All i know: a study in autoepistemic logic. *Artificial Intelligence Journal*, 42:381–386, 1990.

[Liu04] F. Liu. Dynamic variations: Update and revision for diverse agents. Master's thesis, ILLC, University of Amsterdam, 2004.

[Liu06] F. Liu. Preference change and information processing. Working paper, ILLC, University of Amsterdam, 2006.

[Pla89] J. A. Plaza. Logics of public announcements. In *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*, 1989.

[Rot06] H. Rott. Shifting priorities: Simple representations for 27 iterated theory change operators. In H. Langerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 359–384. Uppsala Philosophical Studies 53, 2006.

[Sny04] J. Snyder. Product update for agents with bounded memory. Manuscript, Department of Philosophy, Stanford University, 2004.

[Spo88] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper et al., editor, *Causation in Decision, Belief Change and Statistics II*, pages 105–134. Kluwer, Dordrecht, 1988.

[Was00] R. Wassermann. *Resource Bounded Belief Revision*. PhD thesis, ILLC University of Amsterdam, 2000.

Partial Planning for Situated Agents based on Active Logic

Sławomir Nowaczyk

Slawomir.Nowaczyk@cs.lth.se
Department of Computer Science
Lund University, Sweden

Abstract. This paper presents an investigation of rational agents that have limited computational resources and intentionally interact with their environments. We present an example logical formalism, based on Active Logic and Situation Calculus, that can be employed in order to satisfy the requirements arising due to being situated in a dynamic universe. We analyse how such agents can combine, in a time-aware fashion, inductive learning from experience and deductive reasoning using domain knowledge. In particular, we consider how partial plans are created and reasoned about, focusing on what new information can be provided as a result of action execution.

1 Introduction

In our research we are interested in building rational agents that can interact with their environment. In order to be practically useful, such agents should be modelled as having bounded computational resources. Moreover, since they are situated in a dynamic world, they need to be aware of the notion of time — in particular, that their reasoning process is not instantaneous. On the other hand, such agents have the possibility to acquire important knowledge by observing the environment surrounding them and by analysing their past interactions with it.

This paper mainly focuses on presentation of one kind of logical formalism we think may be appropriate for such agents. We describe how Active Logic can be augmented with epistemic concepts and combined with mechanism related to Situation Calculus, in order to provide flexible and efficient reasoning formalism for rational agents.

We also present how such agents can deal with planning in domains where complexity makes finding complete solutions intractable. Clearly, it is often not realistic to expect an agent to be able to find a total plan which solves a problem at hand. Therefore, we investigate how an agent can create and reason about *partial plans*. By that we mean plans which bring it somewhat closer to achieving the goal, while still being simple and short enough to be computable in reasonable time. Currently we mainly concentrate on plans which allow an agent to acquire additional knowledge about the world.

By executing such "information-providing" partial plans, an agent can greatly simplify subsequent planning process — it no longer needs to take into account the vast number of possible situations which will be inconsistent with newly observed state of the world. Thus, it can proceed further in a more effective way, by devoting its computational resources to more relevant issues.

If the environment is modelled sufficiently well (for example, if a simulator exists), the agent may have a high degree of freedom in exploring it and in deciding how to interact with it. It may be possible to gain information that the agent would not be able to, by itself, observe directly. In many domains it is significantly easier to build and employ a simulator than to analytically predict results of complex interactions. In other cases, for example when the agent is a robot situated in an unknown environment, it must learn "in the wild" and be aware that the actions it executes are final: they do happen and there is no way of undoing them, other than performing, if possible, a reverse action.

In order to accommodate all of the above we use a variant of Active Logic (1) as agent's reasoning formalism. It was designed for non-omniscient agents and has mechanisms for dealing with uncertain and contradictory knowledge. We believe that Active Logic is a good reasoning technique for versatile agents, in particular as it has been applied successfully to several different problem domains, including some in which planning plays a very prominent role (2). Moreover, in order to be able to intentionally direct its own learning process, the agent needs ability to reason about its own knowledge and lack of thereof — thus, the logic we use has been augmented with epistemic concepts (3).

In other words, our agents are supposed to combine deductive and inductive reasoning with time-awareness. We believe that the interactions among those three aspects are crucial for developing truly intelligent systems. It is not our goal to analyse strict deadlines or precise time measurements (although we do not exclude a possibility of doing that), but rather to express that a rational agent needs the ability to reason about committing its resources to various tasks (4). In particular, it is not justified to assume that an agent knows all deductive consequences of its own beliefs.

2 Wumpus Game

The example problem we will be using through this paper is a well-known game of Wumpus, a classic testbed for intelligent agents. In its basic form, the game takes place on a rectangular board through which an player is allowed to move freely. A beast called Wumpus occupies one, initially unknown, square. Agent's goal is to kill the creature, a task that can be achieved by shooting an arrow on that square. Luckily, Wumpus is a smelly creature, so the player always knows if the monster is nearby. But unfortunately, not the exact direction to it. At the same time, when walking around, the player needs to avoid stumbling across the monster, or else he gets eaten by it.

This game is concise enough to be explained easily, but finding a solution is sufficiently complex to illustrate the issues we want to emphasise. We look at it as one instance of a significantly broader class of problems, along the lines of *General Game Playing*, where an agent accepts a formal description of an arbitrary game and, without further human interaction, can play it effectively.

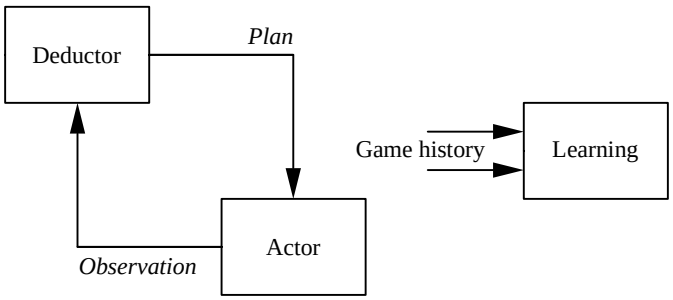


Fig. 1. The architecture of the system.

3 Agent Architecture

We use a simple architecture for our agent, as presented in Fig. 1. It consists of three main elements, corresponding to the three main tasks of the agent.

The Deductor reasons about world, possible actions and what could be their consequences. Its main aim is to generate plans applicable in agent's current situation and predict — at least as far as past experience, imperfect domain knowledge and limited computational resources allow — effects each of them will have, in particular what new knowledge can be acquired.

The Actor is responsible for overseeing the reasoning process, mainly for introducing new observations into the knowledge base and for choosing plans for execution. Basically, it decides *when* to switch from deliberation to acting, and which of the plans under consideration to execute.

These two modules form the core of the agent. By creating and executing a sequence of partial plans our agent moves progressively closer and closer to its goal, until it reaches a point where a winning plan can be directly created by Deductor, and its correctness can be proven.

The learning module is necessary in order to ensure that the plans agent chooses for execution are indeed "good" ones. After the game is over, regardless of whether the agent has won or lost, learning system inductively generalises experience it has gathered — attempting to improve Deductor's and Actor's performance. Our goal is to use the learned information to fill gaps in the domain knowledge, to figure out generally interesting reasoning directions, to discover relevant subgoals and, finally, to more efficiently select the best partial plan.

4 Deductor

In this section we first explain the knowledge representation used, then point out some of the more interesting issues with the reasoning machinery, and finally we introduce a simplified example illustrating how our example Wumpus domain could be axiomatised.

4.1 Knowledge Representation

The language used by Deductor is the First Order Logic augmented with some mechanisms similar to those of Situation Calculus. Within a given situation, knowledge is expressed using standard FOL. In particular, we do not put any limitations on the expressiveness of the language. Predicate $Knows$ describes knowledge of the agent, e.g.,

$$Knows[\ Smell(a) \leftrightarrow \exists_x Wumpus(x) \wedge Neighbour(a, x)\ ]$$

means: *agent knows that it smells on exactly those squares which neighbour Wumpus' position*. The predicate $Knows$ may be nested, a feature which is very useful, if only in a number of specialised contexts — mainly for allowing an agent to reason about future and about what effects performed actions will have. We use standard reification mechanism for putting formulae as parameters of a predicate (5).

In order to describe actions and change, we employ a variant of well-known Situation Calculus approach (6), with an exception that we use predicate $Knows$ instead of $Holds$ — in order to make it explicit that out main priority is describing agent's knowledge. There are some semantic differences between our $Knows$ and a typical $Holds$, however — for example,

$$Knows(s, \alpha \vee \beta) \rightarrow Knows(s, \alpha) \vee Knows(s, \beta)$$

is *not* a tautology, while a corresponding formula for $Holds$ usually is. We will discuss these issues later in this section.

Nevertheless, we introduce an additional parameter to the $Knows$ predicate, which denotes the current situation. Moreover, since the agent is required to reason about knowledge-producing actions, we add yet another parameter, namely the plan agent is going to execute.

In other words, the first formula in this section should be more properly written in the form:

$$Knows[s, p,\ Smell(a) \leftrightarrow \exists_x (Wumpus(x) \wedge Neighbour(a, x))\ ]$$

and mean: *agent knows that executing plan p in situation s leads to a new situation, such that it smells on exactly those squares which neighbour Wumpus' position*. This particular formula is an universal law of the world, valid regardless of the chosen s and p, but many interesting ones — e.g. "$Wumpus(a)$" or "$Knows[Smell(b)]$" — are true only for specific s and p.

The initial knowledge of the agent, the one concerning current situation, represents the state of the world after execution of the empty plan. From this, using typical `STRIPS`-like representation of actions (i.e. preconditions and effects), the agent can reason about extending a particular plan by various operations. Next, for every plan obtained this way, it can deduce which formulae are valid in situation(s) resulting from its execution.

One important extension to the classical `STRIPS` formalism we employ is support for conditional actions. Such actions are important when some of agent's actions may provide information which is necessary to carry on. Basically, the plans agent reasons

about consist of a concatenation of classical and conditional actions, the latter of the form $(predicate\ ?\ action_1\ :\ action_2)$. Those have the standard meaning, i.e. that $action_1$ will be executed if $predicate$ holds, and $action_2$ will be executed otherwise. For a well-developed discussion of other possible ways of representing conditional partial plans and of interleaving planning and execution see, for example, (7).

Conceptually, the agent creates new plans by inductively extending each and every plan considered up to now by each and every possible action. Obviously, many plans created this way would be either invalid or clearly uninteresting. Therefore, due to the computational complexity issues of such a naive approach, we implement plan creation and validation as a process external to the logical reasoner. In general, the main requirement is to ensure that each plan an agent reasons about is valid, i.e. that the preconditions of each action are fulfilled.

We use standard Situation Calculus representation of the actions agent can execute, i.e. using pre– and postconditions. Since some of those actions can be knowledge-producing, it is important to represent that fact properly in their effects. For example, moving onto square a will make an agent know whether $Smell(a)$ is true or false. Of course, this information will only be available during plan execution, not during the planning itself. But it is necessary to distinguish that agent *will have* this knowledge, so that it is able to create a conditional plan branching on value of this predicate. In order to properly represent this notion, we introduce a predicate $KnowsIf$, syntactically similar to $Knows$, except that the meaning of $KnowsIf(s, p, \alpha)$ is that *agent knows that executing plan p in situation s leads to a new situation, in which it* will know *whether α is true or false — but it does not necessarily know, at planning time, which one.*

Extra care needs to be taken when creating conditional plans, as it is important to make sure that the agent has, during plan execution, enough knowledge to correctly choose the appropriate conditional branch. Basically, a precondition for a conditional action $(predicate\ ?\ action_1\ :\ action_2)$ is the conjunction of preconditions for $action_1$, preconditions for $action_2$ and truthfulness of $KnowsIf(predicate)$ — the fact that agent knows whether $predicate$ holds or not.

Finally, due to semantical differences between $Holds$ and $Knows$ predicates, we have found it advantageous to introduce an explicit distinction between *fluents* and *domain constraints*. Semantically, the difference is that truth value of fluents depends on current situation, while the truth value of domain constraints remains fixed for the duration of game episode (although *agent's knowledge* of them can, obviously, change).

To this end we introduce a predicate $Invariant$, distinguishing formulae which are independent of current situation, and a special inference rule which allows agent to propagate knowledge of domain constraints from one situation to another.

4.2 Reasoning

We will start this subsection by describing the reasoning process *within* a given situation, i.e. while an agent ponders which actions to execute. Later we will shortly mention problems that arise when an action is executed and the state of the world changes, and we will conclude with some reference to reasoning about *different game episodes*, which is still very much work in progress.

As we said previously, an agent will create a number of plans and each such plan will be evaluated by an Actor. Therefore, Deductor can devote more effort to plans that are more promising. Since the agent employs Active Logic, this can be easily modelled within that formalism, as it is intended to describe the deduction as an ongoing process, rather than characterising only some static, fixed-point consequence relation.

Active Logic annotates every formula with a time-stamp (usually an integer) of when it was first derived, incrementing the label with every application of an inference rule:

$$\frac{i: \quad a, a \rightarrow b}{i+1: \quad b}$$

It also includes the Now predicate, true only during current time point (i.e., "$i : Now(j)$" is true for all $i = j$, but false for all $i \neq j$). It can, therefore, use this time-stamp to prioritise plans. For example, if an Actor divides the plans into two classes, *normal* ones and *great* ones, we can have inference rules of the kind:

$$\frac{i: \quad Great(s,P), Knows(s,P,a), Knows(s,P,a \rightarrow b)}{i+1: \quad Knows(s,P,b)}$$

$$\frac{i: \quad Now(i), even(i), Knows(s,P,a), Knows(s,P,a \rightarrow b)}{i+1: \quad Knows(s,P,b)}$$

which would ensure that reasoning about *great* plans happens at every step, but reasoning about *normal* ones only happens every other step. This is a very simple example, more complex scenarios are certainly possible, but their usefulness has to be evaluated experimentally. In particular, we find the idea of allowing *the agent* to consciously balance this tradeoff very stimulating.

In a similar spirit, we can balance the tradeoff between creating more complex, longer plans and reasoning about effects of already created plans. At this point it is, however, somewhat unclear on what basis should the agent make decisions regarding this. Nevertheless, another feature of Active Logic, namely the *observation function*, which delivers axioms that are valid since a specific point in time, can be used quite naturally to acquire new plans — possibly from a module external to the reasoner itself.

We mentioned above that the $Invariant$ predicate requires a specialised inference rule. There are several possibilities, one being:

$$\frac{i : Knows(s,p,\alpha) \wedge Invariant(\alpha)}{i+1 : Knows(s',p',\alpha)}$$

for every s, p, s', p'. In practice, the reasoner would not need to multiply the formulae for every possible combination of plan and situation, it can simply mark them as being invariant and take this into account during inference in an efficient way.

After the reasoning and planning has progressed sufficiently, the Actor will choose a plan and execute it. At that point the reasoner can discard whatever other plans it has created — they are no longer suitable — and needs to adapt to the new state of affairs. First, it needs to notice that current situation has changed — we use a predicate

$State$, modelled closely after standard Active Logic predicate Now, in order to achieve this. And second, it needs to absorb the new knowledge acquired by the executing the actions, which can be done using observation function.

An interesting issue is also how to allow an agent to reason about past game episodes. In particular, after the game is won or lost and a new one is being started, a lot of knowledge acquired previously still remains valid and interesting. We are working on ways to extract general useful knowledge and to discover similarities in episodes. This is also closely related to the learning module, discussed in more detail in subsequent sections.

4.3 Wumpus Example

In this section we briefly introduce our example domain, the game of Wumpus. We use a rather natural set of axioms to describe it. First we specify that Wumpus is on exactly one square:

$$\exists_x \, Wumpus(x)$$

$$\forall_{x,y} \;\; x \neq y \rightarrow \neg Wumpus(x) \vee \neg Wumpus(y)$$

We define the smelling phenomenon:

$$\forall_x \, Smell(x) \leftrightarrow \exists_y \, Wumpus(y) \wedge Neighbour(x, y)$$

and that the agent will always know whether it smells on the square it is on:

$$\forall_x \, Player(x) \leftrightarrow KnowsIf(Smell(x))$$

We also define $Neighbour$ relation and other geometrical knowledge in a natural way.

Finally, we need to specify that $Wumpus$ remains stationary thorough the episode. This is the main reason for introduction of the $Invariant$ predicate in previous section. It allows us to state in a simple way that predicates such as $Wumpus$, $Neighbour$, up, $down$, $left$, $right$, $even$ etc. do not depend on the current situation. Therefore, any formula α containing only those predicates is an invariant. So, if an agent discovers in some situation s_1 that $Wumpus(a) \vee Wumpus(b)$, it can easily deduce that $Wumpus(a) \vee Wumpus(b)$ also holds in any other situation s_2.

Observe that a, somewhat more natural, rule such as:

$$\forall_{x,s,s',p,p'} Knows(s, p, Wumpus(x)) \leftrightarrow Knows(s', p', Wumpus(x))$$

does not quite work, as agent's knowledge is typically of the kind

$$Knows(s, p, Wumpus(a) \vee Wumpus(b))$$

and it is not clear how to propagate such formulae between situations without exponential blowup of axioms (problems arise, for example, from the fact that $Knows$ is not distributive to disjunction).

We currently perform grounding of all the variables in order to have reasoning in predicate logic.

4.4 Summary

In general, our agent reasons on three distinct levels of abstraction. First is *within* a given situation, where it tries to find the best plan to execute. Second level concerns effects of executed actions, where state of the world changes and new knowledge becomes available. And third deals with comparing different game episodes, where knowledge previously assumed to be $Invariant$ is no longer so (for example, Wumpus may now be at another location).

The Active Logic formalism used, especially predicate Now, makes it possible to reason about passing time and, combined with observation function delivering knowledge about external events, allows the agent to remain responsive during its deliberations. This way both Deductor and Actor can keep track of how the reasoning is progressing and make informed decisions about balancing thinking and acting.

One of the reasons we have chosen symbolic representation of plans, as opposed to a policy (an assignment of value to each state–action pair) is that we intend to deal with other types of goals than just reachability ones. For a discussion of possibilities and rationalisation of why such goals are interesting, see for example (8), where authors present a solution for planning with goals described in Computational Tree Logic. This formalism allows to express goals of the kind "value of *a* will never be changed", "*a* will be eventually restored to its original value" or "value of *a*, after time *t*, will always be *b*" etc.

To summarise, our agent uses Active Logic to reason about its own knowledge, which is very important in the Wumpus domain. Here, the main goal can be reduced to "learn the position of Wumpus", so active planning for knowledge acquisition is crucial. Agent also requires an ability to compare what kind of information will execution of each plan provide, in order to be able to choose the best one of them.

5 Actor

The Actor module supervises the deduction process and breaks it at selected moments, e.g., when it notices a particularly interesting plan or when it decides that sufficiently long time has been spent on planning. It then *evaluates* existing partial plans and executes the best one of them. The evaluation process is crucial here, and we expect the subsequent learning process to greatly contribute to its improvement. In the beginning, the choice may be done at random, or some simple heuristic may be used. After execution of partial plan, a new situation is reached and the Actor lets the Deductor create another set of possible plans.

This is repeated as many times as needed, until the game episode is either won or lost. Losing the game clearly identifies bad choices on the part of the Actor and leads to an update of the evaluation function.

Winning the game also yields feedback that may be used for improving this function, but it also provides a possibility to (re)construct a complete plan, i.e. one which starts in the initial situation and ends in a winning state. If such a plan can be found, it may be subsequently used to quickly solve any problem instance for which it is applicable. Moreover, even if such plan is not directly applicable, an Actor can use it

when evaluating other plans found by the Deductor. Those with structure similar to the successful one are more likely to be worthwhile.

6 Learning

When analysing learning module, it is important to keep in mind that our agent has a dual aim, akin to the exploration and exploitation dilemma in reinforcement learning. On one hand, it wants to win the current game episode, but at the same time it needs to learn as much general knowledge as possible, in order to improve its future performance.

Currently we are mainly investigating the learning module from the perspective of the Actor — using ILP to evaluate quality of partial plans is, to the best of our knowledge, a novel idea. One issue is that work on ILP has been dealing almost exclusively with the problem of *classification*, while our situation requires *evaluation*. There is no predefined set of classes into which plans should be assigned. What our agent needs is a way to choose the *best* one of them.

For now, however, we focus on distinguishing a special class of "bad" plans, namely ones that lead to losing the game. Clearly some plans — those that in agent's experience *did* so — are bad ones. But not every plan which does not cause the agent to lose is a *good* plan. Further, not every plan that leads to *winning* a game is a good one. An agent might have executed a dangerous plan and win only because it has been lucky.

Therefore, we define as positive examples those plans which lead, or can be proven to *possibly* lead, to the defeat. On the other hand, those plans which can be proven to *never* cause defeat are negative examples. There is a third class of plans, when neither of the above assertions can be proven. We are working on how to use such examples in learning most effectively.

Nevertheless, this is only the beginning. After all, in many situations a more "proactive" approach than simple *not-losing* is required. One promising idea is to explore the epistemic quality of plans: an agent should pursue those which provide the most important knowledge. Another way of expressing distinction between good and bad partial plans, one we feel can give very good results, is discovering relevant subgoals and landmarks, as in (9).

7 Environment Interaction

Another interesting issue in our framework is the "consciousness" of interactions between an agent and its environment, conducted in such a way as to maximise the knowledge that can be obtained. In particular, an agent is facing, at all times, the exploration versus exploitation dilemma, i.e., it both needs to gather new knowledge *and* to win the current game episode.

In order to facilitate such reasoning, our agent requires an ability to both act in the world and to observe it. Finally, it needs to consider its own knowledge and how it will (or *can*) change in response to various events taking place in the environment. In different domains and applications different models of interactions with the world are possible.

The most unrestrictive case is a simulator, where an agent has complete control over the (training) environment. It can setup an arbitrary situation, execute some actions and observe the results. Such a scenario is common in, for example, a physical modelling, where it is often much easier to simulate things than to predict their behaviour and interactions. In a similar spirit, it may be easier for our agent to "ask the environment" about validity of some formula than to prove it.

If agent's freedom is slightly more restricted, it is possible that it is not allowed to freely change the environment, but can "try out" several plans in a given situation. For example, the agent may provide a set of plans and receive an outcome for each of them. Alternatively, it may store some opaque *situation identifier* so that it can revisit the same situation at later time. This model is also suitable for agents that do not have perfect knowledge of the world, as the "replay" capability does not assume *the agent* is able to fully reconstruct the situation or knows the state of the world completely.

In our opinion, this is the most interesting setting: it gives the agent sufficient freedom to allow it to achieve interesting results and at the same time is not, in many domains, overly infeasible. On the other hand, we are working on ways in which this setting could be made even more practical — one idea is having an agent accept the fact that in several replays "the same" situation could vary slightly. For example, physical agent might request an operator to restore the previous state of the world: it would not really be identical, but it may be sufficiently close. Alternatively, in some application domains, only a subset of situations may be "replayable" — only those, for example, that an agent can restore, with required tolerance, all by itself.

In most applications, however, the agent is only able to influence its own actions and have no control whatsoever over the rest of the world. This is also the most suitable model for an *autonomous* physical agent. In such case, the environment will irreversibly move into the subsequent state upon each agent's action (or any other event), leaving it no option but to adapt. It may still be interesting, in some situations, to substitute acting for reasoning, but the agent needs to be aware that once acted upon, the current situation will be gone, possibly forever. It thus needs to consider if saving some deduction effort is indeed the best possible course of action, or if doing something else instead would be more advantageous.

Finally, we can imagine a physical agent situated in a *dangerous* environment, where it is not even plausible for it to freely choose its actions — it needs to, first, assert that an action is reasonably safe. In this case, unlike the previous one, a significant amount of reasoning *needs* to be performed before every experiment.

As an orthogonal issue, sometimes it is feasible for an agent to execute an action, observe the results, reason about them and figure out the next action to perform. But in many applications the "value" of time varies significantly. There are situations where an agent may freely spend its time meditating, and there are situations where decisions must be made quickly. For example, in RoboCup robotic soccer domain, when the ball is in possession of a friendly player, the agent just needs to position itself in a good way for a possible pass — a task which is not too demanding and leaves agent free to ponder more "philosophical" issues. On the other hand, when the ball is rolling in agent's direction, time is of essence and an agent better had plans ready for several most plausible action outcomes.

8 Related Work

Combination of planning and learning is an area of active research, in addition to the extensive amount of work being done separately in those respective fields.

There has been significant amount of work done in learning about what actions to take in a particular situation. One notable example is (10), where author showed important theoretical results about PAC-learnability of action strategies in various models. In (11) author discussed a more practical approach to learning Event Calculus programs using Theory Completion. He used extraction-case abduction and the ALECTO system in order to simultaneously learn two mutually related predicates ($Initiates$ and $Terminates$) from positive-only observations. Recently, (12) developed a system which is able to learn low-level actions and plans from goal hierarchies and action examples provided by experts, within the SOAR architecture.

The work mentioned above focuses primarily on learning how to act, without focusing on reaching conclusions in a deductive way. In a sense, the results are somewhat more similar to the reactive-like behaviour than to classical planning system, with important similarities to the reinforcement learning and related techniques.

One attempt to escape the trap of large search space has been presented in (13), where relational abstractions are used to substantially reduce cardinality of search space. Still, this new space is subjected to reinforcement learning, not to a symbolic planning system. A conceptually similar idea, but where relational representation is actually being learned via behaviour cloning techniques, is presented in (14).

Recently, (15) showed several ideas about how to learn interesting facts about the world, as opposed to learning a description of a predefined concept. A somewhat similar result, more specifically related to planning, has been presented in (16), where the system learns domain-dependent control knowledge beneficial in planning tasks.

Yet another track of research focuses on (deductive) planning, taking into account incompleteness of agent's knowledge and uncertainty about the world. Conditional plans, generalised policies, conformant plans, universal plans and some others are the terms used by various researchers (17; 7) to denote in principle the same idea: generating a plan which is "prepared" for all possible reactions of the environment. This approach has much in common with control theory, as observed in (18) or earlier in (19). We are not aware of any such research that would attempt to integrate learning.

9 Conclusions

The work presented here is still very much in progress and a discussion of an interesting track of research, rather than a report on some concrete results. We have introduced an agent architecture facilitating resource-aware deductive planning interwoven with plan execution and supported by inductive, life-long learning. The particular deduction mechanism used is based on Active Logic, in order to incorporate time-awareness into the reasoning itself. The plans created in deductive way are conditional, accounting for possible results of future actions, in particular information-gathering ones.

We intend to continue this work in several directions. Discovering subgoals and subplans seems to be one of the most useful capabilities of human problem solving and

we would like our agent to invent and use such concept. In our example domain a useful subgoal could be “First, find a place where it smells.” In addition, Deductor should be able to conceive general rules of rational behaviour, such as “Don’t shoot if you don’t know Wumpus’ position”. Yet another clear advantage would be the ability to reuse a previously successful plan in a different situation. Finally, domain experts often are an invaluable source of knowledge that the agent should be able to exploit, if possible.

The ideas for future work mentioned above do not cover all the possible further investigations and extensions of the proposed system; they are just a biased presentation of the authors’ own interests and judgements.

Bibliography

[1] Elgot-Drapkin, J., Kraus, S., Miller, M., Nirkhe, M., Perlis, D.: Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-4072, University of Maryland (1999)

[2] Purang, K., Purushothaman, D., Traum, D., Andersen, C., Perlis, D.: Practical reasoning and plan execution with active logic. In: Proceedings of the IJCAI-99 Workshop on Practical Reasoning and Rationality. (1999) 30–38

[3] Fagin, R., Halpern, J.Y., Vardi, M.Y., Moses, Y.: Reasoning about knowledge. MIT Press (1995)

[4] Chong, W., O'Donovan-Anderson, M., Okamoto, Y., Perlis, D.: Seven days in the life of a robotic agent. In: GSFC/JPL Workshop on Radical Agent Concepts. (2002)

[5] Reiter, R.: Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. The MIT Press (2001)

[6] McCarthy, J., Hayes, P.J.: Some philosophical problems from the standpoint of artificial intelligence. Machine Intelligence **4** (1969) 463–502

[7] Bertoli, P., Cimatti, A., Traverso, P.: Interleaving execution and planning for nondeterministic, partially observable domains. In: European Conference on Artificial Intelligence. (2004) 657–661

[8] Bertoli, P., Cimatti, A., Pistore, M., Traverso, P.: A framework for planning with extended goals under partial observability. In: International Conference on Automated Planning and Scheduling. (2003) 215–225

[9] Hoffmann, J., Porteous, J., Sebastia, L.: Ordered landmarks in planning. Journal of Artificial Intelligence Research **22** (2004) 215–278

[10] Khardon, R.: Learning to take actions. Machine Learning **35** (1999) 57–90

[11] Moyle, S.: Using theory completion to learn a robot navigation control program. In: 12th International Conference on Inductive Logic Programming. (2002)

[12] Könik, T., Laird, J.: Learning goal hierarchies from structured observations and expert annotations. In: 14th International Conference on Inductive Logic Programming. (2004)

[13] Džeroski, S., Raedt, L.D., Driessens, K.: Relational reinforcement learning. Machine Learning **43**(1/2) (2001) 7–52

[14] Morales, E.P.: Relational state abstraction for reinforcement learning. In: Proceedings of the ICML'04 Workshop on Relational Reinforcement Learning. (2004)

[15] Colton, S., Muggleton, S.: ILP for mathematical discovery. In: 13th International Conference on Inductive Logic Programming. (2003)

[16] Fern, A., Yoon, S., Givan, R.: Learning domain-specific control knowledge from random walks. In: International Conference on Automated Planning and Scheduling. (2004)

[17] Cimatti, A., Roveri, M., Bertoli, P.: Conformant planning via symbolic model checking and heuristic search. Artificial Intelligence **159**(1-2) (2004) 127–206

[18] Bonet, B., Geffner, H.: Planning and control in artificial intelligence: A unifying perspective. Applied Intelligence **14**(3) (2001) 237–252

[19] Dean, T., Wellman, M.P.: Planning and Control. Morgan Kaufmann (1991)

Commitment-Based Decision Making for Bounded Agents

Olivier Roy*

Institute for Logic, Language and Computation
Plantage Muidergracht 24
1018TV Amsterdam
The Netherlands
oroy@science.uva.nl

Abstract. This paper explores ways to introduce commitment, as conceived in terms of plans and intentions, into decision theory. An intention-based representation theorem is proved, along with a proposal to model how plans simplify decision making for resource-bounded agents.

1 Introduction

In this paper I adapt the decision theoretical framework proposed in [Anscombe and Aumann, 1963] (AA) to represent formally rational decision-making by agents who are *committed* to certain plans of action. In Section (2) I present the decision theoretical formalism, and Section (3) I sketch the theory of planning agency I'm relying on. In Section (4), I envisage two ways of introducing plans in the formalism. First, I propose to make explicit how plans help resource-bounded agent to simplify decision-making. Second, I show that one can easily use plans to build an *intention-based* utility ranking that is independent of the usual "preference-based" ranking. To facilitate the reading, all the proofs have been moved to the Appendix.

This is not the first attempt at formalizing plan-based decision making. [Bratman et al., 1991] is probably one of the best known. Beside being much closer to decision theory, the framework I propose accounts for one function of plans (modeled in Section 4.2) that is left out in their approach. [van Hees, 2003] also proposed something similar to what I do in Section 4.3, but in a totally qualitative framework. Although I won't explore them here, I think that there are interesting connections to be made between the present work and [McClennen, 1990].

2 Decision making under uncertainty

Decision theory is often called the study of *rational* decision making under *risk* or *uncertainty*. Typically, "uncertainty" is thought of as *subjective*, that is, referring to the partiality of the decision-maker's information. Whereas "risk" is *objective* that is, about the outcome of certain random events such as a lottery, "uncertainty" is *subjective* that is, relative to the decision maker's partial information. In the decision theoretical framework I will use, both risk and uncertainty are modeled. I rely heavily on the presentation of [Myerson, 1991, chap.1], but the framework goes back at least to [Anscombe and Aumann, 1963]. Of course, this is neither the only nor the simplest decision theoretical approach available. See for example [Luce and Raiffa, 1957], where a more classical model is introduced, and further (early) references can be

* I am grateful to Richard Bradley, the LRBA referees, Johan van Benthem, Mikaël Cozic and the attendees at the PALMYR-2 workshop held in Paris in October 2006 for their useful comments.

found. I have chosen to work within the AA framework because it models both random happenings (objective probabilities) and uncertainties (subjective probabilities).

Given an arbitrary set A, I will use $\Delta(A)$ to denote any probability distribution on A. A *decision problem*, $\mathcal{DP}$, is a tuple $\langle X, \Omega, L, \Xi, Pref \rangle$ such that:

- X and Ω are both finite sets of *prizes* and *states*, respectively. I will use x, y, z to range over elements of X and $t, t', t_1, t_2...$ to range over elements of Ω.
- $L = \{f : \Omega \rightarrow \Delta(X)\}$ is the set of all *lotteries* over the prizes in X. $f(x|t)$ should intuitively be read as "the probability of getting prize x given that the current state is t". $[x]$ denote the lottery that gives x for sure, in all states. That is, $[x] = f$ such that $f(x|t) = 1$ for all t. These lotteries are intended to represent random happenings *in the world*. They are "objective uncertainties" or "roulette lotteries" in AA terminology. These contrast with "horse lotteries" or "subjective uncertainties", which represent the agent's beliefs about the true state of the world. In the AA framework, the firsts are sharply distinguished from the seconds. Subjective uncertainties are captured by $\Xi = \{S | S \subseteq \Omega \ \& \ S \neq \varnothing\}$, the set of events.
- $Pref : \Xi \rightarrow \mathcal{P}(L \times L)$ is a function that gives a weak linear preference ordering on lotteries, given an event in Ξ. I will follow Myerson and write $f \preceq_S g$ to say that g is at least as good as f, given that the true state of the world is in S. $f \sim_S g$ and $f \prec_S g$ are defined as usual.

The crux of the AA decision-theoretical framework is to show that, given certain axiomatic constraints, decision problems can be represented by a quantitative utility function and a probability distributions where the preferred lotteries are exactly those that maximize expected utility. More precisely, given a decision problem $\mathcal{DP}$, a *conditional-probability function* $p : \Xi \rightarrow \Delta(\Omega)$ is a function that gives the probability of a state t given that an event $S \in \Xi$ occurs. This will be the probabilistic representative of the agent's subjective uncertainty. We constraint p as follows, for any t and S:

$$p(t|S) = 0 \text{ if } t \notin S \text{ and } \sum_{r \in S} p(r|S) = 1$$

Now, an *utility function* is any function $u : X \times \Omega \rightarrow \mathbb{R}$. This function is used to compute the *expected utility value* of a lottery f given an even S, $E_p(u(f)|S)$, as follows:

$$E_p(u(f)|S) = \sum_{t \in S} p(t|S) \sum_{x \in X} u(x,t) f(x|t)$$

With these definitions in hand, the representation theorem is proven via the method of [Savage, 1954], where the agent's probabilistic beliefs and utility are extracted from his (event-based)-preferences over objective lotteries. Below are Myerson's axioms and statement of the representation theorem. See [Myerson, 1991, p.14-17] for details of the proof.

1. (a) *(Completeness)* $f \preceq_S g$ or $g \preceq_S f$.
   (b) *(Transitivity)* If $f \preceq_S g$ and $g \preceq_S h$ then $f \preceq_S h$.
2. *(Relevance)* If, for all $t \in S$, $f(\bullet|t) = g(\bullet|t)$ then $f \sim_S g$.
3. *(Monotonicity)* If $f \preceq_S g$ and $0 \leq \beta < \alpha \leq 1$ then $(1-\beta)f + \beta g \prec_S (1-\alpha)f + \alpha g$.

4. *(Continuity)* If $h \preceq_S g$ and $g \preceq_S f$ then there exists a number γ such that $0 \leq \gamma \leq 1$ and $g \sim_S \gamma f + (1-\gamma)h$.
5. (a) *(Weak Objective substitution)* If $f \preceq_S e$, $h \preceq_S g$ and $0 \leq \alpha \leq 1$, then $\alpha f + (1-\alpha)h \preceq_S \alpha e + (1-\alpha)g$.
   (b) *(Strict Objective substitution)* If $f \prec_S e$, $h \preceq_S g$ and $0 < \alpha \leq 1$, then $\alpha f + (1-\alpha)h \prec_S \alpha e + (1-\alpha)g$.
6. (a) *(Weak Subjective substitution)* If $g \preceq_S f$, $g \preceq_T f$ and $S \cap T = \varnothing$ then $g \preceq_{S\cup T} f$.
   (b) *(Strict Subjective substitution)* If $g \prec_S f$, $g \prec_T f$ and $S \cap T = \varnothing$ then $g \prec_{S\cup T} f$.
7. *(Non-triviality)* For all $t \in \Omega$ there exist $y, z \in X$ such that $[z] \prec_{\{t\}} [y]$

Theorem 1. *Given a decision problem $\mathcal{DP}$, the following are equivalent:*

1. *$\mathcal{DP}$ satisfies the axioms enumerated above.*
2. *There exists a utility function u and a conditional probability function p such that :*
   (a) For all $t \in \Omega$, $\max_{x\in X} u(x,t) = 1$ and $\min_{x\in X} u(x,t) = 0$.
   (b) For all R, S, T such that $R \subseteq S \subseteq T \subseteq \Omega$ and $S \neq \varnothing$, $p(R/T) = p(R/S)p(S/T)$.
   (c) For all $f, g \in L$ and $S \in \Xi$, $g \preceq_S f$ iff $E_p(u(g)/S) \preceq E_p(u(f)/S)$.

Now, suppose a decision maker is somehow committed to act in a certain way. It is highly debatable whether such commitments can be straightforwardly represented in this framework, for example by changing the preferences or utility assignments. What is more, if one thinks this *can't* be done, it is not clear how to adapt the framework to capture the notion of commitment. Of course, all this depends on what is meant by "the agent is committed to x". In the next section, I will present one understanding of commitment, based on the notion of *plans of action*, and try to explain why one may want to introduce them as an independent feature in the AA framework, especially when one has resource-bounded agents in mind.

3 Plans of Action and Committed Agency

In a nutshell, the driving idea runs as follows: for agents with limited time and computational capacities, being able to commit to plans of action is useful. Of courses, there are many things an agent can be committed to. Here I will restrict to commitment in terms of previously adopted intentions and plans. Moreover, the content of these plans and intentions, what an agent is committed to, will be identified with the agent's *goals*. So, by saying that "the agent is committed to x" I will just mean "he has the intention to get x", "he aims for x" or "x is his goal". Granted, this blurs the relationship between plans and goals, as well as other forms of commitments[1]. I see the present proposal as a first step, upon which a more conceptually fine-grained analysis could be based.

The idea that plans of action are useful for bounded agents has been strongly advocated by [Bratman, 1987]. Since my goal here is not to argue for Bratman's views, but rather to propose a way to implement it into decision theory, its main insights will be uncritically reviewed here, with an outright focus on the different functions of intentions and plans.

[1] Although Section (4.3) is much less bound to this restriction of commitment to intentions.

There is a pertaining distinction in action theory between *future-directed* intentions and intentions *in action*. The latter intentions are thought as the "mental components" of actions (see e.g. [O'Shaughnessy, 1973]) that provide "on the fly" guidance toward their own accomplishment. As their name suggests, future-directed intentions are rather about middle- and long-term goals, and are generally viewed as the output of "beforehand" deliberation. Bratman's work is concerned with future-directed intentions, and so will be mine. Thus, for the remaining of this paper, "intention" should just be understood as future-directed.

According to Bratman intentions are mental states that provide a relative *stability* of goals and *commit* agent to their achievement. Like preferences or desires, intentions can be *reason* for action. But Bratman has argued that intentions and desires are irreducible to one another and thus that the belief-desire view on rational choice is at best an incomplete account of decision-making. In this paper, I will *assume* irreducibility of intentions and ask how they can be introduced in decision theory.

Plans, are viewed as hierarchically structured but typically incomplete sets of intentions. "Hierarchically structured" means that a plan contains general intentions upon which more specific intentions are subordinated. However sharp, the most precise intention of a plan needs not to specify in every detail how it shall be carried out. This is why plans are typically incomplete, a feature which is arguably crucial for agents with limited time and computational resources.

Plans are constrained by norms of rationality. They are first required to be *intentions*- as well as *beliefs*-consistent. The intentions they are made of should not contradict each other, and neither should they contradict the agent's beliefs. In Bratman's view[2], intention consistency is grounded in the fact that intentions are *agglomerative*: if one intends that x and intends that y, he must intend that x and y. Note that this distinguishes intentions from desires: we can have contradictory desires but not contradictory intentions, and desires are clearly not agglomerative. Rational plans also call for "means-end coherence": if one intend to x he should also, at some point, come to intend some necessary means to x. Given that plans are typically incomplete, this means that plans "drive" agents towards deliberating on how they will achieve them.

So, in this model, plans of action are not only outputs but also *input* of deliberation. It is as inputs, so Bratman's theory goes, that they help resource-bounded agents to simplify decision problems. Suppose an agent with a plan $\mathfrak{P}$ is facing a certain decision problem. The pressure for intention consistency will *rule out of consideration* options that preclude the achievement of $\mathfrak{P}$. On the other hand, means-end coherence will call for adopting one of the options that promote the achievement of $\mathfrak{P}$. But it does more than that, given that deliberation is itself an activity that takes time and energy. For resource-bounded agents, it is likely that pondering endlessly over small details of the available options will itself prove to be an obstacle to the achievement of $\mathfrak{P}$. The other way around: careful management of the time and energy devoted to deliberation might prove to be, in itself, an important mean to achieve one's goals. As such, it doesn't seem unreasonable to suppose that means-end coherence will *fix the level of detail up to which options are going to be considered*, thus avoiding useless mind boggling. These two norms, intention-consistency and means-end coherence, are thus likely to simplify deliberation, an advantageous prospect for limited agents.

To sum up, Bratman proposes a "build up" view of planning agency and practical reasoning. Suppose that at some point an agent form a plan $\mathfrak{P}$ made of the intention i_1 to obtain the goal X later. As some

[2] A view that is also endorsed by other action theorists, such [Velleman, 2006] and [Wallace, 2003]

opportunities show up he will have to choose among various attainments of $\mathfrak{P}$, select one of them, say X', form the new intention i_2 to get X' and add it to $\mathfrak{P}$. $\mathfrak{P}$ will be thus enriched until it is achieved[3]. In the next sections, I will consider, in turn, how to model the effects of intention-consistency and means-end coherence on rational *plan-based* decision making under uncertainty. To my knowledge, such proposal is an original contribution. As a side issue, I will also show how one can use plans and intention to, so to speak, set a decision agenda that isindependent of the preferences. As far as this paper goes, all these models will be *static*. I leave for further work the task of modeling the dynamics of plan update.

4 Decision making with commitments

4.1 Plans of action: a formal picture

In the formalization of utility theory sketched in section (2), the object of choices are *lotteries* but, ultimately, what the agents aim for are the *prizes*. Consequently, it seems reasonable to restrict the content of commitments to prizes. That is, the formal picture I am going to depict is one where agents form intentions to get some prizes, given that a certain event S occurs. Given the view of intentions endorsed here, the intended prizes are *goals* the agent will work for. Note that in the AA framework, prizes are exclusive outcomes, so to speak. In the end, the agent gets *one* of the prizes in X. When I say that the subset X' of X is a goal for the agent, I mean that he will act in order to get *one of them*.

In the following definitions, intentions are assimilated with their content and so plans are viewed simply as sets of sets of prizes. So, given an event S, a plan $\mathfrak{P}_S$ is any set of intentions, i.e. of subsets of X. Now, belief and intention-consistency already impose constraints on what is to be considered a *rational* plan. First, belief-consistency precludes impossible intentions, which are just empty sets of prizes. So we are to require that, for any event S, $\varnothing \notin \mathfrak{P}_S$. Intention-consistency requires that the intentions that constitute a plan don't exclude each other. This will boil down to require that any two intentions in a plan have a non-empty intersection. But recall that, for Bratman, intention-consistency in grounded in the agglomerativity of intention, a property that has a clear formal counterpart: $\mathfrak{P}_S$ will be required to be closed under intersection. It should be clear that intention-consistency follows from the requirements that $\varnothing \notin \mathfrak{P}_S$ and that $\mathfrak{P}_S$ is closed under intersection,

Note that, because X is finite, we automatically get that $(\bigcap_{X \in \mathfrak{P}_S} X) \in \mathfrak{P}_S$. This will correspond to the most precise intention of the plan. For most of the results presented below, we can consider without loss of generality that any plan $\mathfrak{P}_S$ contains only this minimal element, and we will use a special notation for it: $\downarrow\mathfrak{P}_S$. A plan is incomplete if $\downarrow\mathfrak{P}_S$ is not a singleton. When the context makes it obvious, we will omit the event subscript S.

Note also that these simple set-theoretical requirement gives already a kind of hierarchical structure to the plan. Indeed, rational plans are semi-lattices with respect to the inclusion relation, with $\downarrow\mathfrak{P}_S$ the smallest element.

[3] This is indeed a very coarse picture of Bratman's theory of practical reasoning. Many important issues are ignored. Among them are the crucial difference between deciding upon an attainment and adopting an intention - as illustrated by the famous "terror bomber" example - and the subtleties of plan reconsideration. These issues will be ignored in the proposed formalization also. They are thoroughly explored in [Bratman, 1987].

4.2 Consistency and coherence to simplify decision problems

As mentioned above, it seems that some decision problems are just too complex to be handled by agents with limited computational and representational capacities. In the usual decision theoretical models I introduced above, it can even be argued that it is simply impossible to do so. The object of choices being the lotteries, the actual choice set is the whole *uncountable* lottery space L. So, for limited agent, we can assume that, somehow, they don't consider all their options, and the norms of intention-consistency and means-end coherence provide an explanation of *why* it is so.

Intention consistency and cleaning of decision problems. Recall that a plan is intention-consistent if it doesn't contain intentions that exclude one another's achievement. This norm has a clear static consequence on plans viewed as sets of prizes, as we just saw. But it also shapes further deliberations by ruling out, in advance, options that would make the plan inconsistent if they were chosen.

Here I propose a way to model this effect of intentions on deliberation. The decision theoretical framework that I sketched above gives us some latitude on which options are going to be considered inconsistent. To keep the model as general as possible, I will just set that inconsistent lotteries are those that give too little chance of getting a prize in the intended set $\downarrow\mathfrak{P}_S$. In other words, the agent is going to make a preference-based decision among the lotteries that give at least a certain probability p to get a prize in $\downarrow\mathfrak{P}_S$. The value of p will stay undefined, making room for the whole spectrum of attitudes towards the prospect of getting an intended prize. For example, an agent for which $p = 1$ will restrict his choices to those who will guarantee a prize in $\downarrow\mathfrak{P}_S$. At the other extreme, an agent for which $p = 0$ will just decide as a regular utility maximizer, since no lotteries will be excluded from the original choice set. Needless to say that the size of the decision problem will decrease as p increases.

Precisely, we have that, given $0 \leq p \leq 1$, the *p-cleaned* version of a decision problem $\mathcal{DP}$ will contain the same prize (X), states (Ω) and events (Ξ) sets as before. What changes is the choice set, the lottery space. For each event S, define $L'_S = \{f : \forall t \in S, \sum_{x \in \downarrow\mathfrak{P}_S} f(x|t) \geq p\}$, that is, L'_S is the set of lotteries p-consistent with the plan adopted at S. Note that for all event S, L'_S is always non-empty because L is the set of *all* probability distribution over X. The preference relations over these cleaned lottery sets are just the restrictions of the original event-based preference orderings: $\preceq'_S = \preceq_S \restriction L'_S$

With this in hand, the next step is to see whether we can still model agents that decide on p-cleaned decision problems as expected utility maximizers. But the very idea of cleaning is to *remove* lotteries from the choice set, which means that L'_s won't satisfy *continuity* anymore, except in the trivial case where $p = 0$. I won't provide a representation theorem that copes with this difficulty.

One could, of course, change the original preference relations by forming an "indifference cluster" with all lotteries that are cleaned out of the choice set. That is, define a new preference relation $\preceq''_S$ as follows: for all $f, g \in L'_S$, $f \preceq''_S g$ iff $f \preceq_S g$ and for all f, g outside L'_S, $f \sim''_S g$. Furthermore, for all f not in L'_S and g in L'_S, set $f \preceq''_S g$. This would have the immediate consequence of ruling out the lotteries outside L'_S as potential expected utility maximizers, while preserving the preference structure within L'_S. However, this "preference shifting" strategy seems to betray the very motivation of cleaning: reducing the "size" of the choice set L. Clearly, under the preference shift strategy the agent just uses the original choice set. So,

I think that one should avoid such an easy way out if he genuinely aims at modeling the simplification function of plans.

As I said, I will not provide a way out of the failure of continuity for cleaned decision problems. What I'll do instead is simply to point out that certain special cases of preference orderings are representable. Say that a preference relation $\preceq'_S$ is *p-cleaned friendly* only if for all $f \in L$ and $t \in S$, if $\sum_{x \in \downarrow \mathfrak{P}_S} f(x|t) < p$ then there exists a lottery g in L such that $f \sim_S g$ and $\sum_{x \in \downarrow \mathfrak{P}_S} g(x|t) \geq p$. In such cases, we immediately get the following:

Theorem 2. *All p-cleaned friendly preference relations restricted to a p-cleaned decision problem $\mathcal{DP}$ are representable by a utility function u' and a probability distribution p', defined in that same way as in Section 2.*

Means-end coherence and the clustering of decision problems In the previous section, I've proposed a framework to reduce the size of a decision problem by "cleaning out" lotteries that give too low a prospect of getting an intended prize. The idea behind the cleaning procedure is that the chosen options shouldn't go against the achievement of the intentions the agent already has, which would violate intention-consistency. Now I turn to the norm of means-end coherence.

As we saw, this norm calls for adopting effective means to reach intended ends. Put into our framework, this means that an agent will have to choose one particular ways to achieve its plans, one *attainment*. But I also mentioned that, for resource-bounded agents, means-end coherence has a further effect. It presses for a careful management of the time devoted to deliberation, and so requires that the agent leaves irrelevant details out. Again, this can be modeled in the present decision-theoretical framework, by *clustering* lotteries that are the equivalent modulo certain attainments of the plan.

Let a pair $\langle \mathcal{A}, p \rangle$ be a set of *p-attainments* in S for a plan $\mathfrak{P}_S$ where $\mathcal{A}$ is a partition of $\downarrow \mathfrak{P}_S$ and p is such that $1/2 < p \leq 1$. The intuitive idea here is that our agent has to choose between certain mutually exclusive ways to achieve its most precise intention $\downarrow \mathfrak{P}_S$, each of them being one "cell" A_i in the partition $\mathcal{A}$. This partition can be more or less fine-grained, making room for various level of details. For now, I make no assumption regarding *how* this a particular partition of $\downarrow \mathfrak{P}$ came to be considered as *the* set of attainments for $\downarrow \mathfrak{P}_S$. I just take it as given, and see how an agent can use it to reduce the size of his decision problem. The parameter p can be seen as the bound under with the agent considers that a lottery f is giving too little chances to get a prize in one the attainment A_i. I require that p is strictly greater than $1/2$ to make sure that the pair $\langle \mathcal{A}, p \rangle$ will finitely partition L.

This is done as follows. Given a cell $A_i \in \mathcal{A}$ and a set of lotteries L, define the *cluster* C_i of L corresponding to A_i as $C_i = \{f : \forall t, \sum_{x \in A_i} f(x|t) \geq p\}$. A lottery cluster C_i puts together all the lotteries that give a price in A_i with at least probability p. Now, the set of clusters C_i will not completely partition L. We need two more clusters to complete the partition. First, the *p-spread cluster* C_p is defined as $\{f : \neg \exists A_i \text{ s.t. } f \in C_i \text{ but } \sum_{x \in \downarrow \mathfrak{P}_S} f(x|t) \geq p\}$. This cluster will contain the lotteries that, for each cell A_i taken individually, don't give high enough probability to be in the cluster C_i, but that nevertheless reach the required probability p over the *whole* set $\downarrow \mathfrak{P}_S$. Finally, the *out of attainment cluster* $C_\varnothing$ is defined as $\{f : \neg \exists A_i \ s.t. \ f \in C_i \text{ and } f \notin C_p\}$. That is, this cluster contains the lotteries that just don't give high

enough probability to get an intended prize. Observe that this cluster will contain exactly the lotteries that would be p-cleaned, according to the procedure described in the previous section. Given all these clusters, we automatically obtain the following fact:

Fact 3 *For a decision problem $\mathcal{DP}$ and a plan $\downarrow\mathfrak{P}_S$, the set of cluster C_i corresponding to an attainment set $\langle\mathcal{A}, p\rangle$ together with the the p-spread cluster C_p and the out of attainment cluster $C_\varnothing$* finitely partition L.

Given this finite partition, we can construct the new *p-clustered version* decision problem $\mathcal{DP}$, relative to an attainment set $\langle\mathcal{A}, p\rangle$ is defined as follows. Take X, Ω and Ξ as before. Set, for all $S \in \Xi$, $L'_S = \{C_i\}_{A_i\in\mathcal{A}} \cup C_p \cup C_\varnothing$ and take $\preceq'_S$ as a total and transitive weak ordering on L'_S. In a nutshell, the options an agent faces now are sets of lotteries, within each he get what he considers high enough chances to attain his goals.

Here I impose no *a priori* restriction on the cluster ordering $\preceq'_S$. But one might want that it somehow reflects the underlying preference ordering $\preceq_S$. This requirement boils down to the following: The preference ordering over a clustered decision problem in an event S is said to be *preference-aligned* if for all $f, g \in L$, if $f \in C_i$ and $g \in C_j$ and $E_p(u(f)/S) \leq E_p(u(g)/S)$then $C_i \preceq'_S C_j$.

With this definition in hand, the challenge is to fine easily computable ways to order a clustered decision problem in a preference-aligned way. This will be postponed for further work. For now I will content myself with stating an easy fact, that relates clustering and cleaning. The proof of this fact is direct, given the observation made above that the outside of attainment cluster $C_\varnothing$ contains just the lotteries that would be p-cleaned.

Fact 4 *For all decision problem $\mathcal{DP}$ and $1/2 < p \leq 1$, the p-cleaned version of the p-clustered $\mathcal{DP}$ is the same as the p-clustered version obtained after a p-cleaning of $\mathcal{DP}$.*

4.3 "Intention-based" utility theory

We thus have two ways to model the effects of plans of action on deliberation, cleaning and clustering, each of which is related to one rational constraint on intention. Now I turn a somewhat orthogonal issue: modeling intentions as setting a "decision agenda" that is independent of the one set by the preferences.

There are many reasons why one might want such an independent ordering. A striking one is that intentions are arguably irreducible to desires and preferences, and thus that the "intention-based" and "preference-based" rationality can diverse. [Bratman, 1987] made a strong case in favor or distinguishing intentions from desires, mainly on the basis that the latter are *not* subject to consistency requirements. But [Sen, 2005] also argued that commitment-based rationality fall systematically out of preference-based decisions. Even if Sen's idea of commitment seems to differ in many respects from Bratman's, these two points of view call for incorporating a intention- or commitment-based ranking of options that is independent of preferences. This can be easily done.

Given a decision problem $\mathcal{DP}$ and a plan $\mathfrak{P}$, introduce a new ordering $\preceq'_S$ defined as follows: For all f, g in L and $S \in \Xi$,

$$f \preceq'_S g \text{ iff for all } t \in S, \sum_{x\in|\mathfrak{P}_S} f(x|t) \leq \sum_{x\in|\mathfrak{P}_S} g(x|t)$$

The equivalence and strict versions of $\preceq'_S$ are defined as usual, which means:

$$f \sim'_S g \text{ iff for all } t \in S \sum_{x \in \mathfrak{P}_S} f(x|t) = \sum_{x \in \mathfrak{P}_S} g(x|t)$$

and similarly for $\prec'_S$.

Notice the intention-based ordering is *derived* from the intentions of the agent, while the preference ordering is *primitive*. It turns out that this new ordering can be quite easily represented by a utility function, proviso a certain structure of the plan.

For a given decision problem, the set of event-relative plans $\{\mathfrak{P}_S\}_{s \in \Xi}$ is said to *reflect boolean operations on events* if for all $S, T \in \Xi$,

1. $\mathfrak{P}_{S \cap T} = \mathfrak{P}_S \cap \mathfrak{P}_T$
2. $\mathfrak{P}_{S \cup T} = \mathfrak{P}_S \cup \mathfrak{P}_T$
3. $\mathfrak{P}_{S-T} = \mathfrak{P}_S - \mathfrak{P}_T$

These conditions are needed to make sure that the plan-based preferences satisfy *Subjective Substitution*, a key property to get a representation in the AA framework. They, so to speak, constraint the inter-event behavior of the plan, and at the same time of the induced orderings $\preceq'_S$. Although they look like constraints imposed for purely technical reasons, they can be philosophically motivated. The first intuitively corresponds to the following maxim: "if you believe more, then your intentions shouldn't be less precise". The second condition has the converse reading: "if you believe less, then your intentions shouldn't be more precise". The third is, I think, the more objectionable. It says that your plan in case you believe that S occurs but T doesn't should be the same as the plan that result from removing what you planned in case of T from what you planned in case of S. This is an extremely strong condition that doubtfully applies to *all* agents, and it remains an open question to me whether there is a more plausible constraint that still allows a utility representation.

Let me now introduce formally the intention-based utility notions that will be used below. Given a decision problem $\mathcal{DP}$, a *plan-based utility function* is a function $u^{\mathfrak{P}} : X \times \Omega \to \mathbb{R}$. Assume a conditional probability p as defined in Section (2), the *expected intention-based utility value* $E_p^{\mathfrak{P}}(u(f)|S)$ of a lottery f is calculated as follows

$$E_p^{\mathfrak{P}}(u^{\mathfrak{P}}(f)|S) = \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t)$$

Theorem 5. (Intention-utility representation theorem) *For any decision problem $\mathcal{DP}$, plans $\{\mathfrak{P}_S\}_{s \in \Xi}$ that reflects boolean operations on events and preference relations $\preceq'_S$ based on the latter, there exists a plan-based utility function $u^{\mathfrak{P}}$ and a conditional probability function p such that :*

1. *For all $t \in \Omega$, $\max_{x \in X} u^{\mathfrak{P}}(x,t) = 1$ and $\min_{x \in X} u^{\mathfrak{P}}(x,t) = 0$.*
2. *For all R, S, T such that $R \subseteq S \subseteq T \subseteq \Omega$ and $S \neq \varnothing$, $p(R/T) = p(R/S)p(S/T)$.*
3. *For all $f, g \in L_{\mathfrak{P}_s}$ and $S \in \Xi$, $g \preceq'_S f$ iff $E_p^{\mathfrak{P}}(u^{\mathfrak{P}}(g)/S) \leq E_p^{\mathfrak{P}}(u^{\mathfrak{P}}(f)/S)$.*

So we can model "plan-based" decisions as maximization of a utility scale independent of the preferences. At this point, I should mention that it is not at all consensual whether intentions and commitment should be thus integrated in decision theory (see e.g. [Pettit, 2005]). As I said before, I will not attempt to defend

the present approach against such arguments. In fact, I have argued in [Roy, 2005] that such an independent scale is unnecessary as long as decision and game theory deal with *ideal* agents.

Along the same lines, note that the beliefs represented by the probability distribution p are now independent form the "intention-based" preferences. So, in this framework, intentions don't have to influence the beliefs. The only relation between these two mental states is imposed by reflection of boolean operations on events, and rather goes from beliefs to intentions. This leaves open whether having the intention to get some prizes in X implies a higher degree of belief that one will actually get some prizes in X.

5 Conclusion

In this paper, I proposed a framework for intention-committed agency built upon a decision theoretical model. I have shown how to represent an intention-based decision making which, in short, models an agent who tries to get prizes in a specific subset of the set of all possible prizes. I have also proposed a way to represent two key aspects of planning agency for resource-bounded agents: simplification by exclusion of inconsistent options and simplification by ignoring irrelevant details.

Many open questions remain for further work, among which the most urgent are:

- Compare intention- and preference-based decisions and examine their combination in interactive situations (games).
- Find a general result concerning the existence of a utility representation for cleaned decision problem.
- Explore simple and efficient algorithms to align preferences over clustered decision problems on intentions and/or original preferences.

Finally, I should mention that some authors have expressed strong skepticism regarding the very idea of introducing plans of action as an independent feature in decision theory (see for example [Strotz, 1956]). This paper can be seen as a modest attempt to rise to the challenge.

A Appendix

Proof of Theorem 2. Just use the same procedure as in [Myerson, 1991], except that everywhere a lottery outside of L'_S would have to be used to define u' or p', use the one in L'_S to which the agent is indifferent.

Proof of Fact 3. It is enough to show that $\approxeq$ is well defined and an equivalence relation, for it is clear that, by the finiteness of X, the partition corresponding to $\approxeq$ is then finite.

1. Assume $f \in C_i$ and $g \in C_j$ for $i \neq j$. We have to show that $f \neq g$. The case where one of the cluster is $C_\varnothing$ or C_p is trivial. Now $f \in C_i$ means that for all t $\sum_{x \in A_i} f(x|t) \geq p$, and similarly from g. But since $\langle \mathcal{A}, p \rangle$ partitions X, and $1/2 < p \leq 1$ we know that $\sum_{x \in A_j} f(x|t) < p$, and the same for g and A_i, which is enough to show that $f \neq g$.
2. It is easy to see that $\approxeq$ is reflexive and symmetric. Transitivity follows from the fact that $\approxeq$ is well-defined.

Proof of Theorem 5. I first establish that $\preceq'_S$ satisfies Subjective Substitution.

Lemma 1. *If $\{\mathfrak{P}_S\}_{s\in\Xi}$ reflects boolean operations on events then $\preceq'_S$ satisfies subjective substitution.*

Proof. Assume that $\{\mathfrak{P}_S\}_{s\in\Xi}$ reflects boolean operations on events. Clearly, for all events S, $\preceq'_S$ is transitive and complete. Take two lotteries g and f such that $g \preceq'_S f$ and $g \preceq'_T f$, i.e.

$$\sum_{x\in\mathfrak{P}_S} g(x|t) \leq \sum_{x\in\mathfrak{P}_S} f(x|t)$$

and

$$\sum_{x\in\mathfrak{P}_T} g(x|t') \leq \sum_{x\in\mathfrak{P}_T} f(x|t')$$

for all t and t' in S and T, respectively. Assume also that $S \cap T = \varnothing$. We now have to show that

$$\sum_{x\in\mathfrak{P}_{S\cup T}} g(x|t) \leq \sum_{x\in\mathfrak{P}_{S\cup T}} f(x|t)$$

for all $t \in S \cup T$. Now we know by assumption that $\downarrow\mathfrak{P}_{S\cup T} = \downarrow\mathfrak{P}_S \cup \downarrow\mathfrak{P}_T$. In other words, we have,

$$\sum_{x\in\mathfrak{P}_{S\cup T}} g(x|t) = \sum_{x\in\mathfrak{P}_S} g(x|t) + \sum_{x\in\mathfrak{P}_{T-S}} g(x|t)$$

and similarly for $f(x|t)$. Now assume that this is not the case that $g \preceq'_{S\cup T} f$. That is, by completeness,

$$\sum_{x\in\mathfrak{P}_S} f(x|t) + \sum_{x\in\mathfrak{P}_{T-S}} f(x|t) < \sum_{x\in\mathfrak{P}_S} g(x|t) + \sum_{x\in\mathfrak{P}_{T-S}} g(x|t)$$

This is equal to

$$\sum_{x\in\mathfrak{P}_S} f(x|t) - \sum_{x\in\mathfrak{P}_S} g(x|t) < \sum_{x\in\mathfrak{P}_{T-S}} g(x|t) - \sum_{x\in\mathfrak{P}_{T-S}} f(x|t)$$

Now $g \preceq'_T f$ also decompose into:

$$\sum_{x\in\mathfrak{P}_{T\cap S}} g(x|t) + \sum_{x\in\mathfrak{P}_{T-S}} g(x|t) \leq \sum_{x\in\mathfrak{P}_{T\cap S}} f(x|t) + \sum_{x\in\mathfrak{P}_{T-S}} f(x|t)$$

which is the same as

$$\sum_{x\in\mathfrak{P}_{T-S}} g(x|t) - \sum_{x\in\mathfrak{P}_{T-S}} f(x|t) \leq \sum_{x\in\mathfrak{P}_{T\cap S}} f(x|t) - \sum_{x\in\mathfrak{P}_{T\cap S}} g(x|t)$$

But then, by transitivity, we have

$$\sum_{x\in\mathfrak{P}_S} f(x|t) - \sum_{x\in\mathfrak{P}_S} g(x|t) < \sum_{x\in\mathfrak{P}_{T\cap S}} f(x|t) - \sum_{x\in\mathfrak{P}_{T\cap S}} g(x|t)$$

which is impossible, given that $T \cap S \subseteq S$ and that the right side of the last inequality must be greater than 0.

Now I turn to the proof of the main theorem. Let $a_1(\bullet|t)$ and $a_0(\bullet|t)$ be the lotteries defined exactly as in [Myerson, 1991, chap.1]. They respectively give for sure the best and worst prizes at t, according to the preference ordering. These will be used to build the probability function p. Define a "bet on t", which will be used to construct the probability function p, as follows

$$b_S(x|t) = \begin{cases} a_1 \text{ if } t \in S \\ a_0 \text{ Otherwise} \end{cases}$$

Now construct the conditional probability function such that for every $t \in \Omega$, $p(t|S)$ satisfies

$$b_{\{t\}} \sim_S p(t|S)a_1 + (1 - p(t|S))a_0$$

This is essentially Savages' method to extract conditional beliefs from preferences. It is a standard argument to show that p as the required properties. Details can be found in [Myerson, 1991, chap.1].

As for the plan-based utility function, just define $u^{\mathfrak{P}}(x, t)$, for every $x \in X$ and $t \in \Omega$, as follows:

$$u^{\mathfrak{P}}(x,t) = \begin{cases} 1 & \text{if } x \in \downarrow \mathfrak{P}_S \\ 0 & \text{Otherwise} \end{cases}$$

It should now be clear that, for all t,

$$\sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) = \sum_{x \in \downarrow \mathfrak{P}_{\{t\}}} f(x|t)$$

But then, because $\preceq'_S$ satisfies *Subjective Substitution*, this extends to all S. That is,

$$\sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) = \sum_{x \in \downarrow \mathfrak{P}_S} f(x|t)$$

And from that we directly get

$$f \preceq'_S g \text{ iff } \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) \leq \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) g(x|t)$$

Which completes the proof.

Remark 1. Getting a representation of $\preceq'_S$ in terms of expected utility is much easier than for the normal preferences because I have defined this ordering using the quantitative information already contained in the lotteries. This is way I could bypass the usual "extraction" of the utility function from preferences, and directly define $u^{\mathfrak{P}}$. This reveals that, in the end, $u^{\mathfrak{P}}$ is no more than a "coarse-grained" utility function.

References

[Anscombe and Aumann, 1963] Anscombe, F. J. and Aumann, R. (1963). A definition of subjective probability. *Annals Math. Stat.*, 34:199–205.

[Bratman, 1987] Bratman, M. (1987). *Intentions, Plans and Practical Reasons*. Harvard UP, London.

[Bratman et al., 1991] Bratman, M. E., Israel, D., and Pollack, M. E. (1991). Plans and resource-bounded practical reasoning. In Pollock, J. and Cummins, R., editors, *Philosophy and AI: Essays at the Interface*, pages 7–22. MIT Press.

[Luce and Raiffa, 1957] Luce, D. R. and Raiffa, H. (1957). *Games and Decisions; Introduction and Critical Survey*. Dover Publications, Inc.

[McClennen, 1990] McClennen, E. F. (1990). *Rationality and Dynamic Choice : Foundational Explorations*. Cambridge UP.

[Myerson, 1991] Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Harvard UP, 1997 edition.

[O'Shaughnessy, 1973] O'Shaughnessy, B. (1973). Trying (as the mental "pineal gland"). *The Journal of Philosophy*, 70(13, On Trying and Intending):365–386.

[Pettit, 2005] Pettit, P. (2005). Construing sen on commitment. *Economics and Philosophy*, 21(01):15–32.

[Roy, 2005] Roy, O. (2005). What does game theory have to do with plans? Technical report, ILLC, Prepublication Series, PP-2005-13.

[Savage, 1954] Savage, L. J. (1954). *The Foundations of Statistics*. Dover Publications, Inc., New York.

[Sen, 2005] Sen, A. (2005). Why exactly is commitment important for rationality? *Economics and Philosophy*, 21(01):5–14.

[Strotz, 1956] Strotz, R. H. (1955 - 1956). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180.

[van Hees, 2003] van Hees, M. (2003). Intentions, utility and rationality. http://www.philos.rug.nl/~vanhees/.

[Velleman, 2006] Velleman, D. (2006). What good is a will? Downloaded from the author's website on April 5th 2006.

[Wallace, 2003] Wallace, R. J. (2003). Normativity, commitment, and instrumental reason. *Philosophers' Imprint*, 1(3):1–26.

A Strongly Complete Logic of Dense Time Intervals

Michał Walicki, Marc Bezem, Wojtek Szajnkenig
Department of Informatics, University of Bergen
PB. 7800, N-5020 Bergen, Norway
{michal,bezem}@ii.uib.no

Abstract

We discuss briefly the duality (or rather, complementarity) of system descriptions based on actions and transitions, on the one hand, and states and their changes, on the other. We settle for the latter and present a simple language, for describing state changes, which is parameterized by an arbitrary language for describing properties of the states. The language can be viewed as a simple fragment of step logic, admitting however various extensions by appropriate choices of the underlying logic. Alternatively, it can be seen as a very specific fragment of temporal logic (with a variant of 'until' or 'chop' operator), and is interpreted over dense (possibly continuous) linear time. The reasoning system presented here is sound, as well as strongly complete and decidable (provided that so is the parameter logic for reasoning about a single state). We give the main idea of the completeness proof and suggest a wide range of possible applications (action based descriptions, active logic, bounded agents), which is a simple consequence of the parametric character of both the language and the reasoning system.

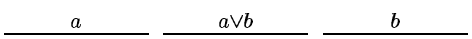
1 Introduction

In the description of reactive systems one has focused primarily on their capability to perform some specific *actions* (process algebras, labelled transition systems, CSP). For example, the famous vending machine can perform the actions of 'accepting a coin' and then 'dispense a coffee' an unspecified number of times. This is certainly a fruitful approach. However, one reason to be interested in actions (and maybe the only one) is that they change the *state* of the world, an agent's beliefs, or any other abstraction. A vending machine can be described equivalently as a device which can stay in a state of 'inactivity', from which it can pass to a state of 'having accepted a coin' and then to one in which it has 'dispensed a coffee'. The latter state may be identified with the state of inactivity if one, among other things, abstracts from the number of coins accepted and from the remaining amount of coffee. These views are in some sense dual, but we present an approach related to the second one, that is, we will describe systems in terms of evolving states. The states evolve over time and during consecutive time intervals certain specifications, that is, partial descriptions of the states, can be observed.

For instance, a description of a system might be as follows. At first an agent knows (or assumes) a. After an announcement he is no longer sure, and knows only $a \vee b$. Finally, after yet another event, he learns that b and retains this knowledge (for the rest of the time: here the scenario ends). This system is described as an expression consisting of a sequence of formulae, each partially describing the state(s), during three consecutive intervals:

$$a; a \vee b; b \tag{1.1}$$

The meaning of this expression could be described on a linear time scale as



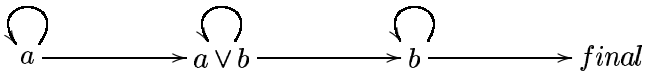
Here, lines represent the 'duration' of the state(s) satisfying the formula which annotates it. One further abstraction is that we view 'duration' as something qualitative but not quantitative. Thus all intervals have been given the same length. Throughout the paper we will depict intervals with different lengths, for convenience, not to suggest different durations.

Viewing the above as a description of the result of some interactions (the announcement and the other event), it is natural to ask for possible consequences, and for an entailment relation between such descriptions in general. For example, a system

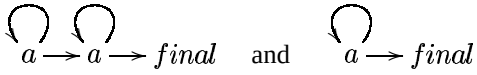
$$a \vee \neg b; a \vee b \tag{1.2}$$

can be viewed as a consequence of the previous one, with the last two intervals concatenated. Furthermore, $a \wedge b$ during a certain interval has both (1.1) and (1.2) as consequence by appropriately cutting the interval in three and two smaller ones, respectively.

Although the paper takes a logical approach, it is possible to interpret the systems above with the help of labelled transition systems. As a consequence of the focus on states instead of on actions, there is just one label. For example, the transition system corresponding to (1.1) is



In our setting, 'being in a state satisfying a specification' is assumed to last for some time, whereas the transitions are instantaneous. The loops in the transition system express this assumption that considered intervals can always be split into smaller ones, that is, the density of the linear ordering modelling the time domain. Transitivity enables the concatenation of intervals. Traces should include the states and are therefore taken to be terminating reduction sequences. Systems are equivalent if they have a trace in common. Equivalent systems have actually infinitely many traces in common, but the trace sets may be different. For example, $a; a$ and a, represented respectively as



are equivalent, but have different trace sets. The consequence relation between systems can also be expressed in terms of transition systems, see section 4.1.

The paper provides the answer to the main question about such systems and their descriptions: given a description (specification) of the sequence of states an agent (or a group of agents) is required to pass through, like (1.1), what other decriptions will be passed through in all cases? In other words, given a language of finite sequences of state-formulae, we ask for an axiomatization which would be strongly complete with respect to intervals of total, dense orderings. Our logic is parameterized by an arbitrary underlying logic of state-formulae. For the sake of illustration we will use propositional logic. Typical examples of possible applications can be obtained from epistemic contexts, like communication protocols or, generally, interacting agents. Due to space limitations, we do not give any more detailed examples, but sketch some possible applications – with particular emphasis on bounded agents – in the last section.

2 Language and Semantics

The logic is parameterized by an arbitrary underlying logic, u.l., which one might want to use for describing the single states. Hence, the language is parameterized by the language of the underlying logic, U.

Definition 2.1 (Sequence Language SL) *The language SL – containing sequence formulae over a parameter language U – is given by the following grammar:*

$$\sigma ::= U \mid \top \mid \sigma; \sigma$$

$\top$ stands for tautology of the underlying logic – this symbol is added only if it is missing in the underlying logic.

In the sequel $\sigma, \sigma_1, \ldots$ denote (sequence) formulae of SL; $f, f_1 ... g, g_1 ...$ – atomic formulae, i.e., those without ; occurring inside. The formulae of our logic will be simple sequents, i.e., have the form $\sigma_1 \Vdash \sigma_2$. We will denote sequents using $q, q', \ldots$. Complexity of a sequence formulae/sequent refers to the number of $\_;\_$ occurring in it.

2.1 Semantics

The semantics is parameterized by the semantics of u.l. Sequence formulae are evaluated over a total, dense ordering which is left-closed and right-open. Given such an ordering $\mathcal{O} = (O, <)$, its points are (mapped to) models of the u.l. An SL-structure is a function

$$r : O \to Mod(u.l.)$$

Left-closedness models the "beginning" (of a computation, or its part), and left-openness its possibly unbounded character. In general, we will consider also subintervals of a whole order $\mathcal{O}$. We denote by $[a, b)$ a *left-closed right-open interval*, $[a, b) = \{o \in O : a \leq o < b\}$, of a given order $\mathcal{O}$. We do not consider empty intervals at all, so the notation $[a, b)$ always implicitly means $a < b$. The satisfaction relation is defined, in general, for any such interval.

Definition 2.2 (Satisfaction Relation) *Satisfaction of an SL-formula σ in an SL-structure, written $[a, b) \models_{\mathcal{O},r} \sigma$, is defined as follows:*

1. $[a, b) \models_{\mathcal{O},r} \top$ *for all* $[a, b)$
2. $[a, b) \models_{\mathcal{O},r} f \iff \forall o \in O : a \leq o < b \Rightarrow r(o) \models_{u.l.} f \quad (f \in u.l.)$
3. $[a, b) \models_{\mathcal{O},r} \sigma_1; \sigma_2 \iff \exists o \in O : a < o < b \,\&$
   $[a, o) \models_{\mathcal{O},r} \sigma_1 \,\&\, [o, b) \models_{\mathcal{O},r} \sigma_2$

We skip the subscripts in the notation, assuming always given O and r. In fact, we will concentrate on the case where the interval is actually the whole O, writing $r \models \sigma$. This is justified by the following equivalence between (2.1) and (2.2). For a semantical entailment, we write $\sigma_1 \models_{\mathcal{O},r} \sigma_2$ iff

$$\forall \mathcal{O} \forall r \forall [a, b) : [a, b) \models_{\mathcal{O},r} \sigma_1 \Rightarrow [a, b) \models_{\mathcal{O},r} \sigma_2 \tag{2.1}$$

Equivalently, we can consider only whole orderings (and not all subintervals):

$$\forall \mathcal{O} \forall r : O \models_{\mathcal{O},r} \sigma_1 \Rightarrow O \models_{\mathcal{O},r} \sigma_2 \tag{2.2}$$

It is an easy exercise to verify that $\_;\_$ is associative. One could view this operator as the until, **U**, of temporal logic over linear time (depending, however, on the details of the definition which may vary). Then, our language could be viewed as a very special subset of temporal logic, where a sequence formula

$$f_1; f_2; f_3; ...; f_n \quad \text{corresponds to} \quad f_1 \mathbf{U} \Big(f_2 \mathbf{U} (f_3 \mathbf{U} ... (f_{n-1} \mathbf{U} \Box f_n) ...) \Big),$$

where the final f_n (and only it) appears always and only under $\Box$, to remain true from then on.[1] Thus it is not surprising that we can express several temporal modalities, for instance:

1. $r \models f$ iff f holds always in r
2. $r \models \top; f$ iff f becomes eventually true and holds then forever
3. $r \models f; \top$ iff f holds initially for at least some time
4. $r \models f; \neg f$ iff f holds for some time, after which $\neg f$ holds forever

Example 2 above admits all the same r's as does 1 but, in addition, also all where f holds almost always, i.e., everywhere with the possible exception of some initial interval. Thus, $f \models \top; f$ but $\top; f \not\models f$. Dually, 3 also allows all models of 1 but also ones where f, holding initially, becomes false after some time, so $f \models f; \top$ but $f; \top \not\models f$. In 4, the requirement is for f to actually become false after some time, never to become true again.

This analogy to temporal logic (of dense linear time) concerns the limited semantic interpretation. However, unlike modal logics, in general, we will offer not only sound and complete, but also strongly complete reasoning, which is also decidable. (All these properties obtain relatively to their presence in the underlying logic.)

[1] Alternatively, one can almost identify $\_;\_$ with the chop operator, common in interval logics, [11, 9]. We have, however, only a very limited fragment of such logics.

The semantics is based on points but, nevertheless, it is strongly interval oriented. For the first, it does not include "point-intervals" (as single points are sometimes called in the interval semantics.) More significantly, although the satisfaction relation is defined relatively to satisfaction in single points, $\models_{u.l.}$, a formula satisfied only at a single point is not satisfied by any interval. For instance, consider $[0,2)$, with all $x \in [0,1) \cup (1,2) : r(x) \models a$ while $r(1) \not\models a$. Then $[0,2) \not\models a$ and $[0,2) \not\models \neg a$. There are some subintervals satisfying a, e.g., $[0,1) \models a$, but there is no subinterval of $[0,2)$ satisfying $\neg a$. This is related also to the phenomenon given in the following example.

Example 2.3 *Assume that u.l. contains propositional disjunction. We may have that $[a,b) \models f \vee g$, while for every subinterval $[c,d) \sqsubseteq_I [a,b) : [c,d) \not\models f$ and $[c,d) \not\models g$.* [2] *Take, for instance, $[a,b)$ such that for any $o \in [a,b), r(o) = (f \vee g) \wedge (\neg f \vee \neg g)$ and, in addition, distribute the models of $\neg f$ densely between those of $\neg g$ and vice versa (i.e., $\forall o,p \in [a,b) : o < p \wedge r(o) \models \neg g \wedge r(p) \models \neg g \Rightarrow \exists q : o < q < p \wedge r(q) \models \neg f$, and the corresponding fact when $r(o), r(p) \models \neg f$.) Then $[a,b) \models f \vee g$ but nowhere, i.e., in no subinterval $[c,d) \sqsubseteq_I [a,b)$, we have that $[c,d) \models f$ or $[c,d) \models g$.*

2.2 Cuts

An equivalent definition of satisfaction can be expressed by saying that an interval $[a,b)$ satisfies a sequence formula $f_1;...;f_n$ iff it is possible to cut the interval into n subintervals, each left-closed right-open and each satisfying the corresponding f_i. This is the content of the following definition.

Definition 2.4 *An n-*cut *C of $[a,b)$ is a partition of $[a,b)$ into n intervals*

$$[a,b)|_C = [a,o_1)[o_1,o_2)\ldots[o_i,o_{i+1})\ldots[o_n,b)$$

with $a < o_i < o_{i+1} < b$.

By $r|_\sigma$ we denote a cut of r which verifies σ, i.e., shows that $r \models \sigma$.[3]

Two cuts of an interval can be superimposed on each other yielding a (possibly) more refined cut.

Example 2.5 ($r|_{\sigma_1 \& \sigma_2}$) *Consider two cuts, each verifying $\sigma_1 = f_1; f_2; f_3$, respectively, $\sigma_2 = g_1; g_2; g_3$. A possible situation is the following:*

$r|_{\sigma_1}$ $\quad$ $[\ \frac{f_1}{g_1}\ |\ \frac{f_2}{g_2}\ \frac{f_2}{g_3}\ |\ \frac{f_3}{g_3}\ )$

$r|_{\sigma_2}$

The result of superimposing these two cuts is as shown below:

$r|_{\sigma_1 \& \sigma_2}$: $\quad$ $[\ \frac{f_1}{g_1}\ |\ \frac{f_2}{g_2}\ |\ \frac{f_2}{g_3}\ |\ \frac{f_3}{g_3}\ )$

The following definition formalizes the superposition of two cuts. It will be used only in the situations where each cut verifies a respective sequence formula and, moreover, when we are considering the sequent $\sigma_1 \Vdash \sigma_2$. Hence, although the constructions are symmetric, we mark the asymmetry $\sigma_1 \Vdash \sigma_2$ in the notation. The operation $Paths(\_)$ collects all possible ways of superimposing a cut verifying $\sigma_1 = f_1;...;f_n$ and one verifying $\sigma_2 = g_1;...;g_m$. In other words, whenever an interval satisfies both formulae, the superposition of the two cuts will satisfy some path in $Paths(\sigma_1 \vdash \sigma_2)$, which is defined below. (We also have the opposite implication, see lemma 2.8.)

Definition 2.6 ($Paths()$) *For an arbitrary sequent q, we define $Paths(q)$ by induction on the complexity l of (number of _;_ in) q :*

$l = 0$: *i.e.,* $q = f \Vdash g$ – $Paths(f \Vdash g) := \{[f \vdash g]\}$.

$l > 0$: *a.* $q = f_1; f_2; \ldots; f_n \Vdash g$

$Paths(q) := \{[f_1 \vdash g] - [f_2 \vdash g] - \ldots - [f_n \vdash g]\}$

[2] $[c,d) \sqsubseteq_I [a,b)$ denotes that $[c,d)$ is a *subinterval* of $[a,b)$, i.e., $[c,d) = \{o \in O : a \leq c \leq o < d \leq b\}$.

[3] This is ambiguous, since there may be many different cuts of r all verifying σ.

b. $q = f \Vdash g_1; g_2; \ldots; g_m$
$Paths(q) := \{[f \vdash g_1] - [f \vdash g_2] - \ldots - [f \vdash g_m]\}$

c. $q = f_1; f_2; \ldots; f_n \Vdash g_1; g_2; \ldots; g_m$
i. $Paths(q) := \{[f_1 \vdash g_1]\} - Paths(f_2; \ldots; f_n \Vdash g_2; \ldots; g_m) \cup$
ii. $\{[f_1 \vdash g_1]\} - Paths(f_1; f_2; \ldots; f_n \Vdash g_2; \ldots; g_m) \cup$
iii. $\{[f_1 \vdash g_1]\} - Paths(f_2; \ldots; f_n \Vdash g_1; g_2; \ldots; g_m)$

Point c. exhausts the possible ways of overlapping of subsequent intervals. Starting with f_1 and g_1, we have the three possibilities illustrated in Figure 1, each corresponding to one of the cases listed under c.

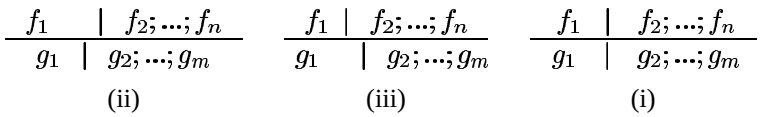


Figure 1: Possible overlapping of initial intervals.

Example 2.7 *$Paths(q)$, for q from example 2.5, are the following:*

$[f_1 \vdash g_1] - [f_2 \vdash g_2] - [f_3 \vdash g_3]$, *(1)*
$[f_1 \vdash g_1] - [f_2 \vdash g_2] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$, *(2)*
$[f_1 \vdash g_1] - [f_2 \vdash g_2] - [f_3 \vdash g_2] - [f_3 \vdash g_3]$, *(3)*
$[f_1 \vdash g_1] - [f_1 \vdash g_2] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$, *(4)*
$[f_1 \vdash g_1] - [f_1 \vdash g_2] - [f_1 \vdash g_3] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$, *(5)*
$[f_1 \vdash g_1] - [f_1 \vdash g_2] - [f_2 \vdash g_2] - [f_3 \vdash g_3]$, *(6)*
$[f_1 \vdash g_1] - [f_1 \vdash g_2] - [f_2 \vdash g_2] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$, *(7)*
$[f_1 \vdash g_1] - [f_1 \vdash g_2] - [f_2 \vdash g_2] - [f_3 \vdash g_2] - [f_3 \vdash g_3]$, *(8)*
$[f_1 \vdash g_1] - [f_2 \vdash g_1] - [f_3 \vdash g_2] - [f_3 \vdash g_3]$, *(9)*
$[f_1 \vdash g_1] - [f_2 \vdash g_1] - [f_2 \vdash g_2] - [f_3 \vdash g_3]$, *(10)*
$[f_1 \vdash g_1] - [f_2 \vdash g_1] - [f_2 \vdash g_2] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$, *(11)*
$[f_1 \vdash g_1] - [f_2 \vdash g_1] - [f_2 \vdash g_2] - [f_3 \vdash g_2] - [f_3 \vdash g_3]$, *(12)*
$[f_1 \vdash g_1] - [f_2 \vdash g_1] - [f_3 \vdash g_1] - [f_3 \vdash g_2] - [f_3 \vdash g_3]$ *(13)*

The path in line (2) corresponds to the cut we obtained in Example 2.5.

A structure r satisfies a path if it satisfies the corresponding sequence formula. For instance, an r satisfies path (2) from the above example iff $r \models f_1 \wedge g_1; f_2 \wedge g_2; f_2 \wedge g_3; f_3 \wedge g_3$.

The following lemma states the observation from example 2.5.

Lemma 2.8 *For all r, σ_1, σ_2 we have that $r \models \sigma_1$ and $r \models \sigma_2$ if and only if there exists a $\pi \in Paths(\sigma_1 \vdash \vdash \sigma_2)$ such that $r \models \pi$.*

3 The reasoning system

Given a sequence formula σ (possibly only a single formula of the underlying logic) we let $\sigma*$ denote σ or an arbitrary extension of σ to a (longer) sequence formula (analogously for $*\sigma$).

Definition 3.1 *The calculus SEC contains the following rule schemata:*

$$\textit{(lift)}\ \frac{}{f \Vdash g}\,[f \vdash g] \qquad (L)\ \frac{f* \Vdash \sigma}{f* \Vdash g; \sigma}\,[f \vdash g]$$

$$(E)\ \frac{\sigma_1 \Vdash \sigma_2}{f; \sigma_1 \Vdash g; \sigma_2}\,[f \vdash g] \qquad (R)\ \frac{\sigma \Vdash g*}{f; \sigma \Vdash g*}\,[f \vdash g]$$

If $*$ is empty, the formula to the left of it remains unchanged. The intuition behind these rules is straightforward and concerns the possible overlapping of subsequent intervals. It refers again to the Figure 1, now with (L) corresponding to (ii), (R) to (iii) and (E) to (i). In either case, validity of the conclusion requires validity of $f \vdash g$ (which is placed in the side-condition). The relation between the remaining parts depends on whether the left formula, f, "lasts longer" than g (L), "shorter than" g (R), or if both have equal duration

(E). Axioms are absent because side-conditions will give proof obligations determining if a given derivation is a proof. The rule (lift) terminates construction of a derivation (bottom-up). A derivation is defined in the standard way. The following gives a couple of examples.

Example 3.2 (Derivation) *Consider the sequent q from Example 2.5. The following are two of its possible derivation:*

$$\Delta_q: \quad \dfrac{\dfrac{\dfrac{}{f_3 \Vdash g_3}\,(lift)\ [f_3 \vdash g_3]}{f_2; f_3 \Vdash g_2; g_3}\,(E)\ [f_2 \vdash g_2]}{f_1; f_2; f_3 \Vdash g_1; g_2; g_3}\,(E)\ [f_1 \vdash g_1] \qquad \Delta'_q: \quad \dfrac{\dfrac{\dfrac{}{f_3 \Vdash g_3}\,(lift)\ [f_3 \vdash g_3]}{f_2; f_3 \Vdash g_3}\,(R)\ [f_2 \vdash g_3]}{f_1; f_2; f_3 \Vdash g_1; g_2; g_3}\,(L)\ [f_2 \vdash g_2]\,(E)\ [f_1 \vdash g_1]$$

We denote by $Der(q)$ the set of all possible derivations of q. Side-conditions in a derivation constitute the proof obligations. Given a derivation Δ_q, we denote by $po(\Delta_q)$ the sequence of all side-conditions (in the bottom-up order). For the derivations from example 3.2, we have

$$po(\Delta_q) = [f_1 \vdash g_1] - [f_2 \vdash g_2] - [f_3 \vdash g_3]$$
$$po(\Delta'_q) = [f_1 \vdash g_1] - [f_2 \vdash g_2] - [f_2 \vdash g_3] - [f_3 \vdash g_3]$$

This operation is extended pointwise to the set of all derivation of a sequent, i.e., $po(Der(q)) = \{po(\Delta) \mid \Delta \in Der(q)\}$.

The following, hardly surprising, lemma states equivalence of the proof obligations obtained over all derivations of a sequent q and the possible overlappings of cuts verifying q.

Lemma 3.3 *For an arbitrary sequent q : $po(Der(q)) = Paths(q)$.*

Definition 3.4 *A derivation Δ_q of $q = f_1; ...; f_n \Vdash g_1; ...; g_m$ is a* proof *of q iff either*

$$\exists 1 \le j \le n: \quad f_j \vdash \bot \text{ (if } \bot \text{ exists in u.l.)} \quad \text{or}$$
$$\forall [f_i \vdash g_j] \in po(\Delta_q): \quad f_i \vdash g_j$$

It is always assumed that, in the underlying logic, $\vdash \top$ and for all $f \vdash \top$.

Example 3.5 *Assuming the underlying logic to be propositional, let the sequent q be as in our previous examples, specialized to actual formulae as follows: $\top; b; \neg b \Vdash \top; \top; \neg b$. Then the derivation Δ_q (from Example 3.2) is a proof of q, but the derivation Δ'_q is not. The latter fails to be a proof, because the side-condition $f_2 \vdash g_3$ is now $b \vdash \neg b$, which does not satisfy the second condition of the definition 3.4.*

Every rule, applied bottom-up, decreases complexity of the sequent. Hence every derivation Δ_q terminates in finitely many steps. Checking if the obtained $po(\Delta_q)$ is a proof is trivially decidable, provided that provability in the underlying logic is decidable. Hence we have a simple, but useful, fact.

Proposition 3.6 *If relation $\vdash$ is decidable, then so is $\Vdash$.*

The next lemma reflects the desired property that we mentioned in the Introduction, i.e., that we do not want to differentiate between f and $f; f$.

Lemma 3.7 *The following rules are admissible.*

$$(\mathrm{id}\Vdash)\ \frac{*f\star \vdash \sigma}{*f; f\star \vdash \sigma} \qquad (\Vdash\mathrm{id})\ \frac{\sigma \Vdash *g\star}{\sigma \Vdash *g; g\star}$$

Thus, we can ignore all adjacent duplicates in the considered sequence formulae. (This fact is used to simplify the proof of lemma 3.3.)

Assuming soundness of the underlying logic, one verifies relatively easily soundness of SEC. Lemma 3.3 is of crucial importance in the proof of completeness, and we now sketch the main steps of this proof.

3.1 Completeness

We assume completeness of the underlying logic. We also restrict our attention to the dense ordering of non-negative rationals, $\mathcal{Q}$, but constructions and the final result generalize to arbitrary dense orderings.

Given an interval $[a,b) \sqsubseteq_I \mathcal{Q}$, and $k \geq 2$ models $c_0, ..., c_{k-1}$, one can distribute them densely, i.e., so that for any two points $a \leq o_1 < o_2 < b$, the subinterval $[o_1, o_2)$ contains all models. We register this fact without proof.

Lemma 3.8 *$k \geq 2$ models $c_0, ..., c_{k-1}$ can be distributed densely over an interval $D = [a,b) \sqsubseteq_I \mathcal{Q}$ so that (the image of) every non-empty subinterval $X \sqsubseteq_I D$ contains all models $c_0, ..., c_{k-1}$.*

Theorem 3.9 *When u.l. is strongly complete, then so is SEC.*

PROOF: We show that every unprovable sequent has a counter-model. So assume a sequent q with no proof: $f_1; \ldots; f_n = \sigma_1 \Vdash \sigma_2 = g_1; \ldots; g_m$ (with no adjacent duplicates). By definition 3.4 combined with lemma 3.3, this means that $\forall \pi \in Paths(q)$:

- $\forall [f_i \vdash g_j] \in \pi : f_i \nvdash \bot$ – all f_i are consistent, AND
- $\exists [f_i \vdash g_j] \in \pi : f_i \nvdash g_j$.

We construct a model of $f_1; ...; f_n$ by constructing an interval $r_i \models f_i$ for every $1 \leq i \leq n$. We use rationals, so for every $i < n$, we let $r_i = [i-1, i)$, while $r_n = [n-1, \infty)$. For every r_i we assign the models as follows.

For every pair f_i, g_j, the proof obligation $f_i \vdash g_j$ occurs on some derivation path. For each f_i, we collect all g_j's (from all derivation paths) such that $f_i \nvdash g_j$. (If for some f_i, there are no such g_j's, then we let r_i contain any model of f_i (existing by $f_i \nvdash \bot$.) We construct r_i as follows:

Since $f_i \nvdash g_j$ (for each chosen g_j) so, by completeness of u.l., $f_i \not\models g_j$, so we have a counter-model, m_{ij}, for each such pair. We collect all such m_{ij} for a given f_i and distribute them densly in the interval r_i. By lemma 3.8, we then have that, for all j for which we have a counter-model $m_{ij} : r_i \not\models g_j$ (and $\forall s \sqsubseteq_I r_i : s \not\models g_j$).

Concatenating all the intervals $r = r_1; r_2; ...; r_n$, we obtain $r \models f_1; ...; f_n$, with the cut $r_i \models f_i$, which we now fix.

If now $r \models g_1; ...; g_m$ (*) then, by Lemma 2.8, there exists a path $\pi \in Paths(\sigma_1 \Vdash \sigma_2)$ such that for every node $[f_i \vdash g_j] \in \pi$, the respective subinterval $r_{ij} \models f_i \wedge g_j$. However, by lemma 3.3, $Paths(q) = po(Der(q))$, i.e., π comprises the proof obligations from one of the derivation paths for q. Since no such path is a proof, it contains a node $[f_i \vdash g_j]$ where $f_i \nvdash g_j$. But then also $\forall s \sqsubseteq_I r_i : s \not\models g_j$ – contradicting (*). □

Example 3.10 *Consider the following (unprovable) sequent $q = \sigma_1 \Vdash \sigma_2$:*

$$a; a \vee b; b \Vdash a \vee b; a; b$$

The $Paths(q)$ are obtained as they were in example 2.7, and are here listed with the formulae f_i, g_j instantiated appropriately:

$[a \vdash a \vee b] - [a \vee b \vdash a] - [b \vdash b]$, *(1)*
$[a \vdash a \vee b] - [a \vee b \vdash a] - [a \vee b \vdash b] - [b \vdash b]$, *(2)*
$[a \vdash a \vee b] - [a \vee b \vdash a] - [b \vdash a] - [b \vdash b]$, *(3)*
$[a \vdash a \vee b] - [a \vdash a] - [a \vee b \vdash b] - [b \vdash b]$, *(4)*
$[a \vdash a \vee b] - [a \vdash a] - [a \vdash b] - [a \vee b \vdash b] - [b \vdash b]$, *(5)*
$[a \vdash a \vee b] - [a \vdash a] - [a \vee b \vdash a] - [b \vdash b]$, *(6)*
$[a \vdash a \vee b] - [a \vdash a] - [a \vee b \vdash a] - [a \vee b \vdash b] - [b \vdash b]$, *(7)*
$[a \vdash a \vee b] - [a \vdash a] - [a \vee b \vdash a] - [b \vdash a] - [b \vdash b]$, *(8)*
$[a \vdash a \vee b] - [a \vee b \vdash a \vee b] - [b \vdash a] - [b \vdash b]$, *(9)*
$[a \vdash a \vee b] - [a \vee b \vdash a \vee b] - [a \vee b \vdash a] - [b \vdash b]$, *(10)*
$[a \vdash a \vee b] - [a \vee b \vdash a \vee b] - [a \vee b \vdash a] - [a \vee b \vdash b] - [b \vdash b]$, *(11)*
$[a \vdash a \vee b] - [a \vee b \vdash a \vee b] - [a \vee b \vdash a] - [b \vdash a] - [b \vdash b]$, *(12)*
$[a \vdash a \vee b] - [a \vee b \vdash a \vee b] - [b \vdash a \vee b] - [b \vdash a] - [b \vdash b]$ *(13)*

On every path there exists a node with unprovable obligation, either $[a \vee b \vdash a]$ or $[a \vee b \vdash b]$ (as

well as $[b \vdash a]$ *or* $[a \vdash b]$*). Hence* $\sigma_1 \not\models \sigma_2$*. The counter-model will be built from three intervals,* $r = [0,1)[1,2)[2,\infty)$*, where* $\forall o \in [0,1) : r(o) = a \wedge \neg b$ *(a boolean structure assigning* `true` *to* a *and* `false` *to* b*),* $\forall o \in [2,\infty) : r(o) = b \wedge \neg a$*, while in* $[1,2)$ *we distribute densly the counter-models for* $a \vee b \vdash b$ *(namely* $a \wedge \neg b$*) and for* $a \vee b \vdash a$ *(namely* $\neg a \wedge b$*). This will ensure that* $[1,2) \models a \vee b$*, and so* $r \models a; a \vee b; b$*, as we have the following situation*

$$r = \begin{array}{c|c|c} r_1 & r_2 & r_3 \\ \hline a \wedge \neg b & a \vee b & \neg a \wedge b \end{array}$$

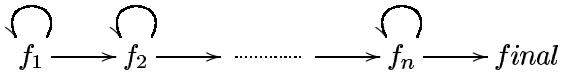
A cut verifying $a \vee b; a; b$ *must first comprise some subinterval verifying* $a \vee b$ *and then some verifying* a*. But the latter can occur only within* r_1*, as* $r_3 \models \neg a$*, while every subinterval* $s \sqsubseteq_I r_2 : s \not\models a$*. But then, the rest of* r *will not satisfy* b*, since all subintervals* $s \sqsubseteq_I r_2 : s \not\models b$*. In short,* $r \not\models a \vee b; a; b$*.*

4 Related systems and applications

The presented system, being interval-based, seems to require a comparison with other interval logics. However, since it possesses a number of desirable properties usually missing in such logics, it promises also potential for applications, and we comment briefly such possibilities. (An overview of other interval logics can be found in [9].) Relations to other frameworks – transition systems, 4.1, action-based descriptions, 4.2, and active logic, 4.3 – indicate also possible applications in the areas where these other techniques are applied.

4.1 Transition systems

As mentioned in the introduction, we can define for every sequence formula $\sigma = f_1; f_2; \ldots; f_n$ a transition system (or: rewrite system):



The *meaning* of σ can now be given as the set of all rewrite sequences of this transition system. Here some care has to be taken. First, we only consider rewrite sequences that end in the final state. Second, we consider the states modulo equivalence in the underlying logic. Third, we adopt some notation of formal language theory and denote the rewrite sequences as words $f_1^{p_1} f_2^{p_2} \ldots f_n^{p_n}$ and call them *traces* (note that we denote only the states, and not the transitions). We use Kleene $^+$ to denote one or more iterations. With these points in mind we define the semantics of σ as follows:

$$[\![\sigma]\!] = f_1^+ \ldots f_n^+$$

For example, $[\![a \wedge b; b \wedge a; c]\!] = (a \wedge b)^+ (b \wedge a)^+ c^+ = \{(a \wedge b)^p c^q \mid p > 1, q > 0\}$. Equivalence of systems can now be defined as follows.

Definition 4.1 $\sigma_1 \cong \sigma_2$ *if* $[\![\sigma_1]\!] \cap [\![\sigma_2]\!]$ *is non-empty.*

Indeed, $\cong$ is an equivalence relation: reflexivity and symmetry are trivial, and transitivity follows after a moment's reflection on the regular expressions involved. The following lemma characterizes the semantic entailment in terms of transition systems. ($\sigma(i)$ denotes the i-th "state" of σ.)

Lemma 4.2 $\sigma_1 \models \sigma_2$ *if and only if*

$$\exists\, \sigma_1' \cong \sigma_1,\ \sigma_2' \cong \sigma_2 : |\sigma_1'| = |\sigma_2'| \ \&\ \forall\, i = 1, \ldots, |\sigma_1'| : \sigma_1'(i) \models_{u.l.} \sigma_2'(i)$$

The correspondence between the traces σ_1' and σ_2' reflects the existence of a joint path $\pi \in Paths(\sigma_1 \Vdash \sigma_2)$ from lemma 2.8, which verifies $\sigma_1 \models \sigma_2$.

4.2 Representing actions

Suppose we want to model an action of sending a message m over a secure channel from (agent) A to B, $send(A,m,B)$. We model the environment as another agent E, and security of the channel means that E cannot see what is communicated between A and B. As the underlying logic, we use here some variant of epistemic logic, where $\mathbf{K}_X(y)$ stands for the statement that agent X knows y (has y available). Communication of m from A to B is modeled by a rewrite rule, which defines it in terms of the effects on the consequtive states:

$$send(A,m,B) \;\leadsto\; s_1 \,;\, s_2$$

Now, for instance, a secure communication corresponds to the fact that, starting in a state where A, but not E, knows m, we pass to a state where both A and B, but still not E, know m. This means that as s_1, s_2 we use:

$$s_1 = \mathbf{K}_A(m) \wedge \neg\mathbf{K}_E(m)$$

$$s_2 = \mathbf{K}_A(m) \wedge \neg\mathbf{K}_E(m) \wedge \mathbf{K}_B(m).$$

A reliable communication, i.e., one which is not only secure, but where A can also be sure that B obtains his message, is modeled simply by adding additional conjunct to the resulting state:

$$s_1 = \mathbf{K}_A(m) \wedge \neg\mathbf{K}_E(m)$$

$$s_2 = \mathbf{K}_A(m) \wedge \neg\mathbf{K}_E(m) \wedge \mathbf{K}_B(m) \wedge \mathbf{K}_A(\mathbf{K}_B(m)).$$

An insecure channel makes it possible for E to intercept the message. Analyzing security aspects, one wants to address the worst case scenario, and therefore lets E actually intercept all messages:

$$s_1 = \mathbf{K}_A(m) \wedge \neg\mathbf{K}_E(m)$$

$$s_2 = \mathbf{K}_A(m) \wedge \mathbf{K}_E(m) \wedge \mathbf{K}_B(m).$$

Once a series of actions is rewritten as a series of their effects, σ_1, we can apply our logic for deriving its consequences, by asking for σ_2 such that $\sigma_1 \vdash \sigma_2$. The point of these examples is to illustrate the flexibility of the proposed setting for handling a virtually unlimited variety of possible action types. This flexibility is achieved by *not* axiomatizing any actions, but merely by representing actions in terms of their effects on the states.

Remark 4.3 *Notation $\mathbf{K}_A(m)$ suggests that we can use some variant of modal epistemic logic, i.e.,* S4 *or* S5*. This is correct, but we should take some precautions in formulating the proof obligations (side-conditions in our reasoning system). For instance, we might want to prove that, given that $a \to b$, an agent A knowing first a, will eventually (be able to) know b, i.e., $\top; \mathbf{K}_A(b)$. The last step in a derivation of such a description would amount to the following:*

$$\frac{}{(a \to b) \wedge \mathbf{K}_A(a) \Vdash \mathbf{K}_A(b)} \;\; [(a \to b) \wedge \mathbf{K}_A(a) \vdash \mathbf{K}_A(b)] \tag{4.1}$$

In modal logic, strong completeness requires that the premisses are included under the $\Box$-modality (here $\mathbf{K}$), and the premise $a \to b$ does not satisfy this condition. The general statement of this fact is as follows (e.g., exercise 1.5.3 in [3]):

$$\Gamma \models^g \phi \iff \mathbf{K}^*(\Gamma) \models^l \phi \tag{4.2}$$

where $\models^g$ is the global logical consequence (which we are using), while $\models^l$ is local logical consequence, for which we have strongly complete reasoning systems.[4] *In the logics where $\mathbf{K}(\phi) \leftrightarrow \mathbf{K}(\mathbf{K}(\phi))$, we can simplify the right hand side of (4.2) to $\Gamma, \mathbf{K}(\Gamma) \models^l \phi$, or even drop Γ when axiom T is present. Consequently, when using modal epistemic logic, like* S4 *or* S5*, we would have to utilize strongly complete versions of the reasoning system, based on (4.2). For proving the side-condition of (4.1), we would obtain a proof ending with the following transition:*

[4] Statement (4.2) can be relativised to arbitrary classes of frames which are closed under 'reachable subframes', i.e., classes M where, when $(W,R) \in \mathsf{M}$ and (W_w, R_w) is a subframe of (W,R) obtained by taking $W_w = \{w' \mid R^*(w,w')\}$ for some fixed $w \in W$ and restricting R to this subset, then also $(W_w, R_w) \in \mathsf{M}$. This closure property is enjoyed by the typical modal logics like S4 or S5.

$$\frac{\vdots}{\dfrac{\mathbf{K}_A(a \to b) \wedge \mathbf{K}_A(a) \vdash \mathbf{K}_A(b)}{(a \to b) \wedge \mathbf{K}_A(a) \vdash \mathbf{K}_A(b)}}$$

4.3 Active logic

Active logic has been developed to reason about the changing states of beliefs, [6, 2]. Although our framework does not aim at equal completeness of modelling, several aspects of active logic fall naturally into it (e.g., allowing for nonmonotonicity, temporally evolving beliefs). Other aspects addressed by active logic have been in our case factored out and delegated to the choice of the underlying logic. Thus, for instance, to model bounded agents, we only have to choose an appropriate logic for such agents; to treat contradictory beliefs, some form of paraconsistent logic could be chosen. We illustrate some of these aspects below.

Active logic operates with the explicit notion of (discrete) time points, and application of rules involves always increase of time. (This simple idea is present in the predecessor of active logic, step logic [4, 5, 7], and in its decidable fragment, [1].) For instance:

$$\frac{\mathsf{n} : f\ ,\ f \to g}{\mathsf{n}+1 : g}$$ – if, at step n, f and $f \to g$ are known/available – then, at step n+1, g can become known/available

Semantical differences notwithstanding, the following rule, admissible in our system, can be used to model exactly this aspect of stepwise reasoning:

$$(step)\ \frac{\sigma \Vdash *f\star}{\sigma \Vdash *f; g\star}\,[f \vdash g]$$

Thus, for instance, the sequence starting with (i) the knowledge that birds fly, $\phi = \forall x(B(x) \to F(x))$, and, after some time, (ii) learning that t is a bird, after which (iii) ϕ, namely, the fact that birds can fly is forgotten, is expressed with the (appropriate fragment of) first-order logic as the underlying logic, as given on the left of $\Vdash$. From this we obtain the proof that the fact $F(t)$ was (possibly) known at some point:

$$\phi; \phi \wedge B(t); B(t) \ \Vdash\ \phi; \phi \wedge B(t); \phi \wedge B(t) \wedge F(t); B(t) \ \Vdash\ \top; F(t); \top$$

The modelling is more abstract than in active logic where one simply counts the applications of rules. This is possible here as well, but we can also allow much coarser granularity admitting transitions to arbitrary consequences. (Of course, counting of steps happens in our case only at the meta-level. Inclusion of the time-step, step-sequence, etc. into the agent language, so central in step and active logics, would require appropriate choice of the underlying logic.)

4.4 Bounded agents

Boundedness of agents, which underlies the stepwise model of reasoning, can be captured in our setting simply by using appropriate (semantics of the) underlying logic. For instance, one might apply the logic of finite agents from [10], with the collection of formulae as its syntactic models of the subsequent states. With this logic one can express that an agent knows a, b, $\mathbf{K}_A(a, b)$, as well as that he knows *at most* b, c, d, written $\overline{\mathbf{K}}_A(b, c, d)$. Then we can prove that an agent who first (i) knows a, then (ii) knows a, b, c and then (iii) at most b, c, d, eventually does not know a:

$$\mathbf{K}_A(a); \mathbf{K}_A(a, b, c); \overline{\mathbf{K}}_A(b, c, d) \Vdash \top; \neg\mathbf{K}_A(a)$$

More intricate and practical examples of applications of the logic presented here, and its combination with the logic for bounded agents, can be found in [14], or obtained by contacting one of the authors.

More generally, one can model bounded agents by distinguishing the explicit (limited) and the implicit (potentially unlimited) knowledge, cf. [13, 8]. Here, we would interpret the left hand side of $\Vdash$ as the description of the actual sequence of states of *explicit* knowledge of an agent (each with a finite model). The derivable right hand sides of $\Vdash$ represent then possible *implicit* consequences of such transitions. For instance, using the (step) rule, we can unfold possible consequences of the following initial state:

$$\mathbf{K}_A(a \vee b, \neg a) \ \Vdash\ \mathbf{K}_A(a \vee b, \neg a); \mathbf{K}_A(a \vee b, a \vee b \vee c, \neg a); \mathbf{K}_A(a \vee b, a \vee b \vee c, \neg a, b)$$

The meaning is that the states on the right of $\vdash$ are reachable from the (initial) state on the left by finitely many deduction steps. With such an interpretation, as the underlying logic one might even choose any variant of modal epistemic logic, and yet avoid the omniscience problems inherent to these logics. This particular application of our logic is, in fact, the approach taken in the logic of algorithmic knowledge, [12].

5 Conclusions and Future Work

We have presented a temporal logic with a single binary modality, roughly corresponding to **U**ntil, and interpreted it over dense linear time. It is parameterized by an *arbitrary* underlying logic for description of the states. Our logic inherits metaproperties of the underlying logic: its is strongly complete/sound/decidable whenever the underlying logic is. We have also suggested the possible way of handling actions in our framework, simply by defining them through their effects on the states of the system. What is missing at the present stage, are detailed examples. The parametric character of our logic offers possibility of applying it to a wide variety of contexts, but the details and usefulness of such applications remain to be investigated.

Also, we would like to adjust the logic to a wider variety of orderings. A simple restriction of the (L) rule yields a sound system for *all* total orderings. Although we expect it to be complete, the proof of the fact is still missing. For discrete orderings, the relation $\models$ is decidable, but the problem of constructing a natural reasoning system remains open. This problem turns out to be surprisingly difficult and is currently under investigation.

References

[1] Natasha Alechina, Brian Logan, and Mark Whitsey. A complete and decidable logic for resource-bounded agents. In *3-rd Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*. ACM Press, 2004.

[2] Michael L. Anderson, Walid Gomaa, John Grant, and Don Perlis. On the reasoning of real-world agents: toward a semantics of active logic. In *7-th Annual Symposium on the Logical Formalization of Commonsense Reasoning*, Corfu, Greece, 2005.

[3] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, 2001.

[4] Jennifer Drapkin and Don Perlis. A preliminary excursion into step-logics. In *SIGART International Symposium on Methodologies for Intelligent Systems*, 1986.

[5] Jennifer Elgot-Drapkin. *Step-logic: reasoning situated in time*. PhD thesis, University of Maryland, 1988.

[6] Jennifer Elgot-Drapkin, Sarit Kraus, Michael Miller, Madhura Nirkhe, and Don Perlis. Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-3680 and UMIACS-TR-99-65, University of Maryland, 1999.

[7] Jennifer Elgot-Drapkin, Michael Miller, and Don Perlis. Memory, reason and time: the step-logic approach. In R. Cummins and J Pollock, editors, *Philosophy and AI: Essays at the Interface*. MIT Press, Cambridge, Mass., 1991.

[8] Ronald Fagin and Joseph Y. Halpern. Belief, awareness, and limited reasoning. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, 1985.

[9] Valentin Goranko, Angelo Montanari, and Guido Sciavicco. A road map of interval temporal logics and duration calculi. *Journal of Applied Non-Classical Logics*, 14(1-2):9–54, 2004.

[10] Thomas Ågotnes. *A Logic of Finite Syntactic Epistemic States*. PhD thesis, Department of Informatics, University of Bergen, 2004.

[11] Joseph Y. Halperm and Yoav Shoham. A propositional modal logic of time intervals. *Journal of the ACM*, 38(4):935–962, 1991.

[12] Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. Algorithmic knowledge. In Ronald Fagin, editor, *5-th Conference on Theoretical Aspects of Reasoning about Knowledge (TARK'94)*. Morgan Kaufman, 1994.

[13] Hector J. Levesque. A logic of implicit and explicit belief. In *National Conference on Artificial Intelligence (AAAI-84)*, pages 198–202, 1984.

[14] Wojciech Szajnkenig. *Sequence Logic*. PhD thesis, Department of Informatics, University of Bergen, [forthcoming].